University of Cyprus
Department of Biological Sciences

FACULTY OF PURE AND APPLIED SCIENCES

DEPARTMENT OF BIOLOGICAL SCIENCES

LABORATORY OF BIOTECHNOLOGY AND MOLECULAR VIROLOGY

PROFESSOR LEONDIOS G. KOSTRIKIS

Genetic Characterization of a Novel HIV-1 Circulating Recombinant Form (CRF91_cpx) and Identification of Three Additional Putative CRFs: Increased Prevalence of URFs and CRFs in the Polyphyletic HIV-1 Infection in Cyprus.

Vasilis Georgiou

A Research-Based Master's Thesis
for the
Degree of Magister Scientiae in Biomedical Sciences

May 17, 2022

ABSTRACT

Background: HIV is characterized by a high degree of genetic variation, as it is evident from numerous genetically distinct subtypes (within the major group M) and circulating recombinant forms (CRFs). The genetic diversity of HIV is attributed to its nature as a fast-replicating virus, coupled with a high mutation rate and the lack of proof-reading ability of reverse transcriptase. The HIV-1 epidemic in Cyprus is characterized as highly *poly*phyletic due to the influx of various HIV-1 group M subtypes from Africa, Europe, and Asia. The coexistence of numerous subtypes in the same geographic area raises the potential for the generation of new recombinant strains. Thus, the investigation of the viral HIV-1 nucleotide sequences uncovered the existence of a novel circulating recombinant form (CRF91_cpx) in Cyprus and raised evidence for the existence of additional three CRFs.

Materials and Methods: The four transmission clusters (clusters 4,5,9 and 16) were identified through phylogenetic analyses of HIV-1 pol region (2253-5250 on HXB2 genome) sequences, derived from consenting HIV-1 positive patients in Cyprus (2017-2021) collected as part of routine surveillance and antiretroviral resistance analyses. The HIV-1 genotypic subtypes were determined using REGA-3.0. Subsequently, a multiple sequence alignment was created using CLUSTALW algorithm, followed by phylogenetic analyses performed on the MEGA-X software for the construction of maximum-likelihood trees (GTR model, 1000 bootstrap replicates). The identification of the transmission clusters was performed on Cluster-Picker software (genetic distance ≤0.045, bootstrap support value ≥70%). Later, near full-length HIV-1 genome (790-8795 on HXB2 genome) sequences were obtained, which were aligned against a reference dataset of all known HIV-1 subtypes and CRFs (RIP alignment 2017 and 2020), and phylogenetic analyses were repeated. Next, similarity plot and bootscan analyses (sliding window of 400 nucleotides overlapped by 40 nucleotides) were performed against a reference set of all HIV-1 group M pure subtypes, CRF02_AG and CRF56_cpx using SimPlot v3.5.1 software for each of the HIV-1 recombinant query sequences belonging to the four new CRF strains. Consequently, subregion confirmatory neighbor-joining tree analyses were performed using MEGA-X software for confirmation of the recombination breakpoints and subtype origin of each fragment (Kimura two-parameter model, 1000 bootstrap replicates, ≥70% bootstrap-support value was considered definitive).

Results: In the maximum-likelihood trees of near full-length genome sequences, the HIV-1 recombinant sequences of each cluster exclusively clustered together, revealing their high genetic similarity and uniqueness. The subtypes of the near full-length genome sequences for cluster 9 were determined as “Rec. of 02_AG, G, J”, cluster 4 as “Rec. of 56_cpx, G” and cluster 5 and 16 as “Rec. of A1, B”. The similarity plot and the bootscan analyses illustrated the same unique mosaic pattern for each of the HIV-1 recombinant clusters, revealing six putative inter-subtype recombination breakpoints for the cluster 9 “Rec. of 02_AG, G, J”, two inter-subtype recombination breakpoints for the cluster 4 “Rec of 56_cpx, G”, and seven inter-subtype recombination breakpoints for the two recombinant clusters 5 and 16 “Rec of A1, B”. The neighbor-joining subregion confirmatory analyses confirmed the subtype origin of each fragment.

Conclusions: Thus, it was concluded that the corresponding query sequences of each of the HIV-1 recombinant clusters have the same and unique mosaic pattern, which consists of fragments of subtypes CRF02_AG, G, J, and a fragment of unknown subtype origin for the cluster 9 the subtypes CRF56_cpx and G for the cluster 4, and the subtypes A1 and B for the clusters 5 and 16. In conclusion, we have characterized the mosaic structure of the new HIV-1 circulating recombinant form, named CRF91_cpx (cluster 9) by the Los Alamos HIV Database according to the standards of HIV nomenclature. Since the identification of the CRF91_cpx, 2 more patient samples have been introduced into the CRF91_cpx transmission cluster, demonstrating an active growth. For the cluster 4, “Rec. of 56_cpx, G” and clusters 5 and 16 “Rec of A1, B”, all the evidence indicate that these recombinant strains are novel CRFs, and as such approval is pending from the Los Alamos HIV Database to name them according to the standards of HIV nomenclature.

DEDICATION

Every challenging work need self-efforts as well as guidance of elders especially those who were very close to our heart. My humble effort I dedicate to my sweet and loving Mother whose affection, love, and encouragement, make me able to get such success and honor.

ACKNOWLEDGEMENTS

The successful completion of both my studies and my dissertation is the result not only of my own effort but also of the understanding and help of the people in my work environment and all the professors and office staff of the Department of Biological Sciences.

As such, seizing the opportunity, I would first like to thank my research advisor Professor Leondio G. Kostriki, for the impartial assistance and trust that he has offered me throughout my undergraduate studies, master's studies, and now with my master's dissertation.

Furthermore, I would like to thank the Ph.D. candidate Cicek Topcu that worked with me side by side for three years and guided me every step of the way in this amazing journey.

Finally, I would like to thank the Ph.D. candidate Chrysostomou Andrea and the research scientist and lab manager Rodosthenous Johana, that implemented a crucial role during my undergraduate studies, master's studies, and my master's dissertation.

COMPOSITION OF THE EXAMINATION COMMITTEE

Thesis Supervisor (Examination Committee coordinator): Professor Leondios G. Kostrikis, Department of Biological Sciences, University of Cyprus

Committee Member: Associate Professor Antonis Kirmizis, Department of Biological Sciences, UCY

Committee Member: Assistant Professor Chrysoula Pitsouli, Department of Biological Sciences, UCY

SEMINAR ANNOUNCEMENT

University of Cyprus
Department of Biological Sciences

Master Research Dissertation in Biomedical Sciences (BIO 830/600)

Student Presentation

Friday, 27 May 2022 at 15:00
Building CTF 01, Room 002, Panepistimioupoli Campus

This seminar is open to the public

Vasilis Georgiou

Thesis Supervisor: Prof. Leondios Kostrikis

"Genetic Characterization of a Novel HIV-1 Circulating Recombinant Form (CRF91_cpx) and Identification of Three Additional Putative CRFs: Increased Prevalence of URFs and CRFs in the Polyphyletic HIV-1 Infection in Cyprus."

HIV is characterized by a high degree of genetic variation, as it is evident from numerous genetically distinct subtypes (within the major group M) and circulating recombinant forms (CRFs). The genetic diversity of HIV is attributed to its nature as a fast-replicating virus, coupled with a high mutation rate and the lack of proof-reading ability of reverse transcriptase. The HIV-1 epidemic in Cyprus is characterized as highly *pol*yphyletic due to the influx of various HIV-1 group M subtypes from Africa, Europe, and Asia. The coexistence of numerous subtypes in the same geographic area raises the potential for the generation of new recombinant strains. Thus, the investigation of the viral HIV-1 nucleotide sequences uncovered the existence of a novel circulating recombinant form (CRF91_cpx) in Cyprus and raised evidence for the existence of additional three circulating recombinant forms (CRFs). The four transmission clusters (clusters 4,5,9 and 16) were identified through phylogenetic analyses of HIV-1 *pol* region (2253-5250 on HXB2 genome) sequences, derived from consenting HIV-1 positive patients in Cyprus (2017-2021) collected as part of routine surveillance and antiretroviral resistance analyses. The HIV-1 genotypic subtypes were determined using REGA-3.0. Subsequently, a multiple sequence alignment was created using CLUSTALW algorithm, followed by phylogenetic analyses performed on the MEGA-X software for the construction of maximum-likelihood trees (GTR model, 1000 bootstrap replicates). The identification of the transmission clusters was performed on Cluster-Picker software (genetic distance ≤0.045, bootstrap support value ≥70%). Later, near full-length HIV-1 genome (790-8795 on HXB2 genome) sequences were obtained, which were aligned against a reference dataset of all known HIV-1 subtypes and CRFs (RIP alignment 2017 and 2020), and phylogenetic analyses were repeated. Next, similarity plot and bootscan analyses (sliding window of 400 nucleotides overlapped by 40 nucleotides) were performed against a reference set of all HIV-1 group M pure subtypes, CRF02_AG and CRF56_cpx using SimPlot v3.5.1 software for each of the HIV-1 recombinant query sequences belonging to the four new CRF strains. Consequently, subregion confirmatory neighbor-joining tree analyses were performed using MEGA-X software for confirmation of the recombination breakpoints and subtype origin of each fragment (Kimura two-parameter model, 1000 bootstrap replicates, ≥70% bootstrap-support value was considered definitive). In the maximum-likelihood trees of near full-length genome sequences, the HIV-1 recombinant sequences of each cluster exclusively clustered together, revealing their high genetic similarity and uniqueness. The subtypes of the near full-length genome sequences for cluster 9 were determined as "Rec. of 02_AG, G, J", cluster 4 as "Rec. of 56_cpx, G" and cluster 5 and 16 as "Rec. of A1, B". The similarity plot and the bootscan analyses illustrated the same unique mosaic pattern for each of the HIV-1 recombinant clusters, revealing six putative inter-subtype recombination breakpoints for the cluster 9 "Rec. of 02_AG, G, J", two inter-subtype recombination breakpoints for the cluster 4 "Rec of 56_cpx, G", and seven inter-subtype recombination breakpoints for the two recombinant clusters 5 and 16 "Rec of B, A1". The neighbor-joining subregion confirmatory analyses confirmed the subtype origin of each fragment. Thus, it was concluded that the corresponding query sequences of each of the HIV-1 recombinant clusters have the same and unique mosaic pattern, which consists of fragments of subtypes CRF02_AG, G, J, and a fragment of unknown subtype origin for the cluster 9 the subtypes CRF56_cpx and G for the cluster 4, and the subtypes A1 and B for the clusters 5 and 16. In conclusion, we have characterized the mosaic structure of the new HIV-1 circulating recombinant form, named CRF91_cpx (cluster 9) by the Los Alamos HIV Database according to the standards of HIV nomenclature. Since the identification of the CRF91_cpx, 2 more patient samples have been introduced into the CRF91_cpx transmission cluster, demonstrating an active growth. For the cluster 4, "Rec. of 56_cpx, G" and clusters 5 and 16 "Rec of A1, B", all the evidence indicate that these recombinant strains are novel CRFs, and as such approval is pending from the Los Alamos HIV Database to name them according to the standards of HIV nomenclature.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 INTRODUCTION

1.1 Discovery and Origin of Human Immunodeficiency Virus HIV

The official start of the epidemic took place in the summer of 1981, when the US Center for Disease Control and Prevention (CDC) made a report indicating a cluster of five men who have sex with men (MSM) suffering from pneumocystis carinii pneumonia (PCP), a rare illness that effect immunodeficient individuals (Rickman 1997). In the same period, other MSM developed a rare skin cancer called Kaposi's sarcoma (Friedman-Kien 1981). This unexpected increase in diseases associated with immune deficiency led the scientific community to research the cause. Thus, in 1983 two independent research groups led by the American Robert Gallo and the French Françoise Barré-Sinoussi, characterized a novel retrovirus that was causing the immune deficiency (Barré-Sinoussi, Chermann et al. 1983, Gallo, Sarin et al. 1983). After extensive research it was proved that the two independent groups characterized the same retrovirus, which was named as Human Immunodeficiency Virus (HIV) three years later in 1986 (Case 1986). At the beginning of the epidemic, as this illness was closely related with the MSM community, it was named as Gay Related Immune Deficiency (GRID) (Altman 1982). Soon, it was determined that the name GRID was not representative because the immunodeficiency syndrome wasn't an illness that only affects the MSM but other groups as well, such as intravenous drug users (IDV). As such the July of 1982 the term Acquired Immune Deficiency Syndrome (AIDS) was introduced (Kher 1982).

There are two main types of HIV, the HIV type 1, and the HIV type 2. The HIV-2 is genetically distinct from the HIV-1 due to a characteristic that makes the virus less prevalent (Seillier-Moiseiwitsch, Margolin et al. 1994, Nyamweya, Hegedus et al. 2013). Both HIV-1 and HIV-2 appear to be derived from the simian immunodeficiency virus (SIV), which comes from primates in West-Central Africa (Figure 1). The virus was transmitted through a process called "zoonotic transmission" in the early 20th century (Gao, Bailes et al. 1999, Sharp, Hahn 2011, Faria, Rambaut et al. 2014). HIV-2 comes from the transmission of the SIVsmm virus to humans. The SIVsmm virus is the SIV virus that infects the sooty mangabey monkeys living in West Africa (Reeves, Doms 2002). The origins of HIV-1 occurred in the Republic of the Congo through the transmission of SIVcpz to humans. The SIVcpz virus is the SIV virus that infects chimpanzees (Gao, Bailes et al. 1999, Reeves, Doms 2002). Regarding the process of zoonotic transmission of the HIV-1, it has occurred at least three times during evolution creating the three main groups of

HIV-1, which are M (major), N (non-M, non-O) and O (outlier) (Gao, Bailes et al. 1999, Sharp, Bailes et al. 2001). Recently a fourth independent HIV-1 zoonotic transmission was discovered leading to the creation of a rare main group called P (O-like). It was estimated that this group was established though the transmission of SIVcpz to gorilla leading to the modification of the virus creating in such a way the SIVgor. Then the zoonotic transmission of the SIVgor to humans lead to the modification of the virus creating the main group P (Van Heuverswyn, Li et al. 2006, Vallari, Holzmayer et al. 2011). The Figure 1 below, through phylogenetic analysis, illustrate the genetic similarity of HIV-1 and HIV-2 with SIVcpz and SIVsmm respectively as well as the similarity of the SIVcpz with the main groups M, N, O, and P of HIV-1.

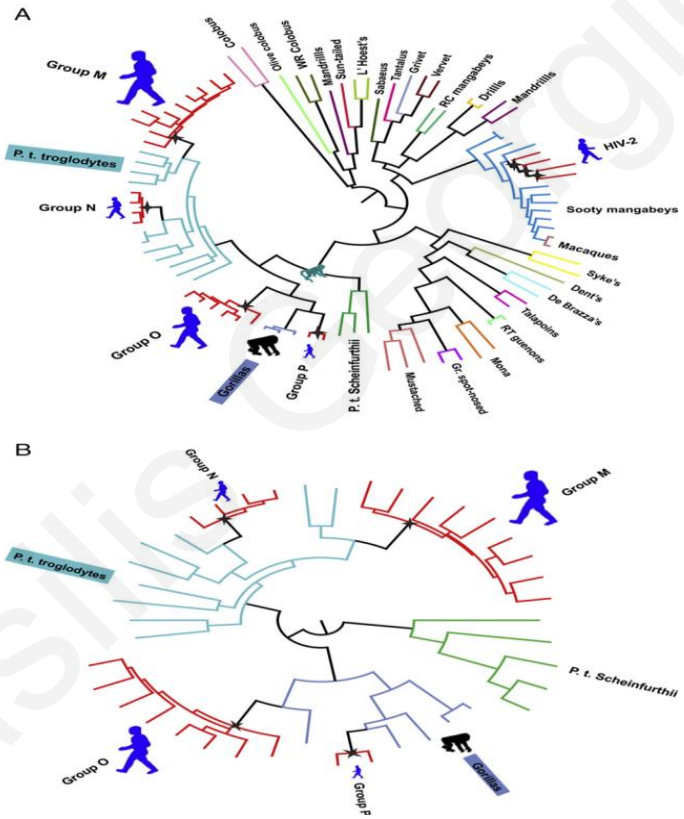


Figure 1: *Phylogenetic tree of the near-full genomic sequence of the virus SIV/HIV.* HIV-1 and HIV-2 are shown in red, while the other colors show the different species of SIV in monkeys. (A) Correlation of HIV-1 and HIV-2 with SIVcpz and SIVsmm, respectively. (B) More detailed correlation of SIVcpz with the main groups M, N, O, and P of HIV-1. The asterisks indicate the zoonotic transmission of the virus (Beloukas, Psarris et al. 2016).

1.2 Taxonomy, Structure and Genome of HIV-1

HIV-1, based on the International Committee on Taxonomy of Viruses (ICTV), is classified in the genus Lentivirus and the family Retroviridae (Lefkowitz, Dempsey et al. 2018). In general, viruses that belong to the Lentivirus genus share common biological characteristics.

The most important ones, among these shared characteristics, are the long-term illness and the long incubation time of the virus (Haase 1986). The virus consists of two positive single-stranded, non-covalently linked RNAs that encode the virus proteins. The virus genome is encapsulated in a conical capsid, with diameter 40-60 nm at the wide end and about 20 nm at the narrow end. The capsule is made up from 2,000 copies of the p24 protein. In addition, the virus genome is strictly bound to nucleoprotein p7. Inside the capsule there are key enzymes, which are necessary for the replication of the virion. These enzymes are the reverse transcriptase (RT), protease (PR), ribonuclease, and integrase (IN). The capsule is surrounded by the matrix, which is made up from the p17 protein. The matrix is surrounded by the viral envelope, that is consist of the lipid bilayer of the host cell, and proteins which are derived from the host as well as from the virus (Turner, Summers 1999). The viral proteins that bind to the envelope is glycoprotein 120 (gp120) and glycoprotein 41 (gp41). The envelope proteins are forming trimers that are consist of three gp41 and three gp120, leading to the formation of a spike in the viral envelope. There are on average 14 spikes in each virion (Clapham, McKnight 2001). The location inside the virion as well as the structure of the aforementioned proteins are represented at the Figure 2 below.

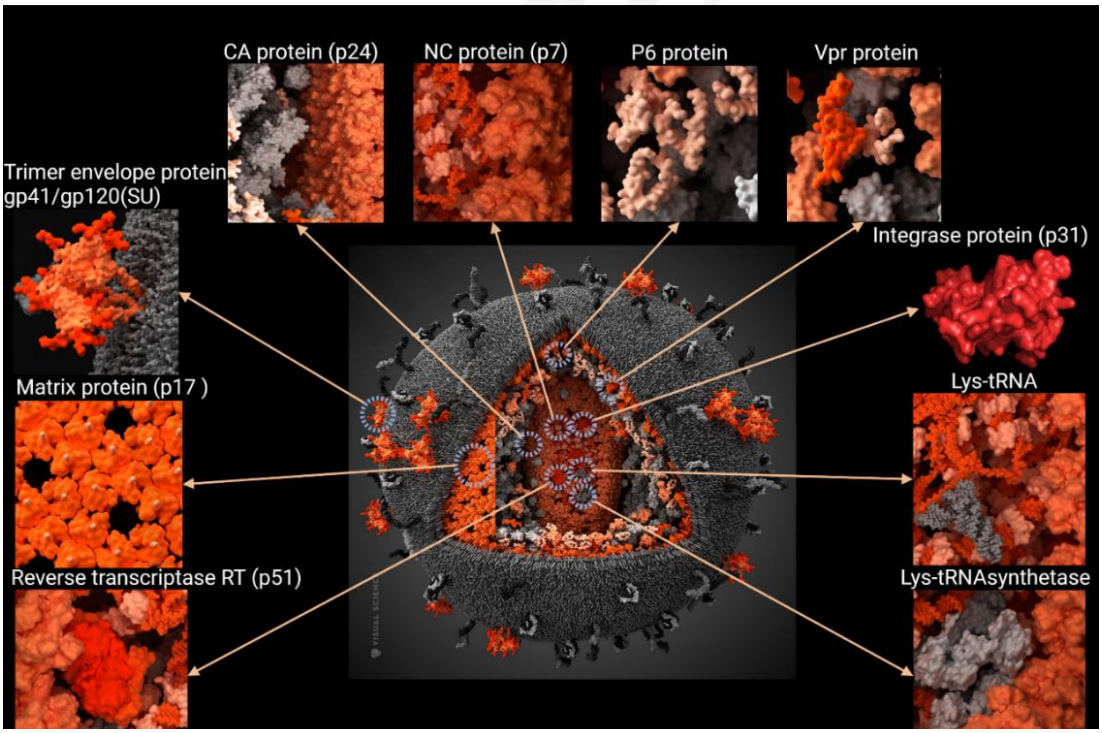


Figure 2: Representation of the structure of the mature HIV-1 virion. The virus particle is surrounded by the structures and names of the proteins that make it constitute. The discontinued blue cycles and the orange arrows indicate the spatial position of these proteins within the virion. The pictures of the figure were taken by visual science(*https://visual-science.com/projects/hiv/illustrations/*), a work that was contacted by (Konstantinov, Stefanov et al. 2010)

The HIV-1 genome is about 9.3 Kb in size. The positive single-stranded, non-covalently linked RNAs, through the RT enzyme, are reverse transcribed to double-stranded DNA. Then, through the integrase protein the viral double stranded DNA is integrated in the DNA of the host cell. Both the 5 'end and the 3' end contain long terminal repeats (LTRs).

The 5 'LTR act as the transcription promoter of the viral genes. In the 5 ' to 3' direction, the gag region encode the outer core membrane protein (MA, p17), capsid protein (CA, p24), nucleocapsid (NC, p7) and a smaller nucleic acid stabilizing protein. This is followed by the pol region encoding the protease (PR, p12), reverse transcriptase (RT, p51) together with RNase H (p15) and integrase (p31). Next to the pol region is the env region, from which the two glycoproteins, gp120 (surface protein) and gp41 (transmembrane protein), are encoded.

In addition to the basic gag, pol and env gene regions, other helper genes are expressed. Those genes perform a variety of roles, for instance, the regulator proteins (Tat, Rev) are necessary for the initiation of HIV-1 replication and the regulatory proteins (Nef, Vif, Vpr, Vpu) have an impact on the viral proliferation, development, and pathogenesis(Blood 2016). The positions as well as the frame of each of the aforementioned genes are represented in the Figure 3.

Moreover, the total of 16 proteins produced from the virus can develop up to 120 possible interactions with either natural proteins of the virus or with functional proteins of the host (Li, G., De Clercq 2016). Consequently, the multi-functionality properties of the virus proteins, the ability to use multiple reading frames of the genome, as well as the ability of the virus proteins to interact with natural proteins of the virus and with the host proteins, give the virus the ability to perform multiple functions with relatively small genome.

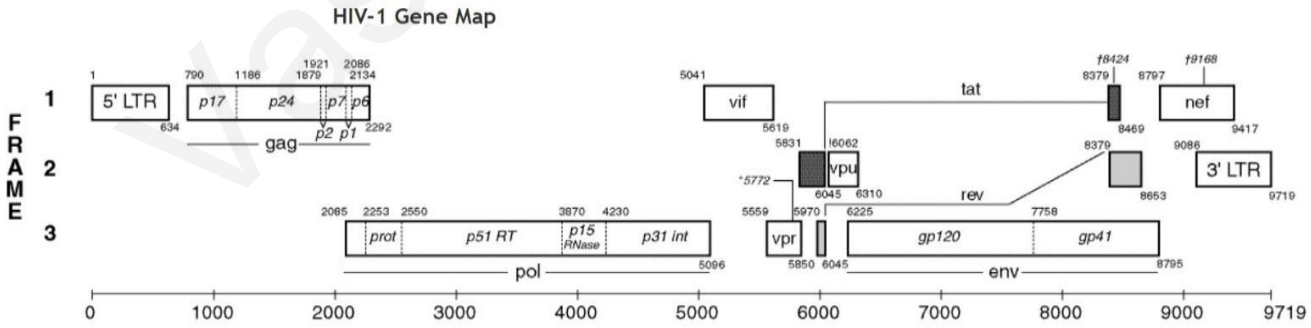


Figure 3: *Schematic representation of the HXB2 reference genome (GenBank accession number*: *K03455*). The open reading frames are shown on the left. For each gene the nucleotide number on the upper left indicates the start and the nucleotide number on the bottom right indicates the end. The numbering was based on the HXB2 reference genome (Landmarks of the HIV genome) (Compendium 2018).

1.3 HIV-1 Tropism

The term "Tropism" refers to the general ability of viruses to infect specific cell types. HIV-1 for its attachment and entry into cells recognizes the CD4 cell receptor through the gp120 protein (Dalgleish, Beverley et al. 1984). To ensure that the binding of the virus to the host receptor is correct, HIV-1 use the gp41 protein to recognize co-receptors as a control. The gp41 protein recognizes chemokine co-receptors, that have the characteristic of the seven hydrophobic transmembrane spanning regions and specifically the co-receptors CCR5 and CXCR4 (Dragic, Litwin et al. 1996, Deng, Liu et al. 1996, Feng, Y., Broder et al. 1996). Therefore, in order for the virus to be able to enter the target cells, they must have the CD4 receptor, and one or both of the CCR5 and CXCR4 co-receptors. It is noteworthy, that not all the strains of the virus recognize both co-receptors, but on the contrary, the minority has this ability. The majority of HIV-1 strains recognize the CCR5 co-receptor and are known as R5 (M-tropic) strains and infect macrophage cells, without rapidly proliferating. The strains of the virus that recognize the CXCR4 co-receptor are known as R4 (T-tropic) strains and infect CD4+ T-Lymphocytes with high infectivity. Additionally, the strains that have the ability to recognize both co-receptors are known as R5X4 strains and can infect both macrophages and CD4+ T-Lymphocytes (Figure 4) (Tersmette, de Goede et al. 1988, Fenyo, Morfeldt-Manson et al. 1988, Fenyo, Albert et al. 1989, Coakley, Petropoulos et al. 2005). With the exception of some rare cases, only R5 and R5X4 viruses are transmitted between people (Wilen, Tilton et al. 2012). It is important to mentioned that some individuals show a deletion of 32 nucleotides in the CCR5 allele (delta32-CCR5) resulting in reduced (heterozygous individuals) or no expression (homozygous atoms) of the CCR5 co-receptor. This mutation makes these individuals extremely resistant to the virus and is found mainly in Europeans (Kostrikis, Huang et al. 1998)

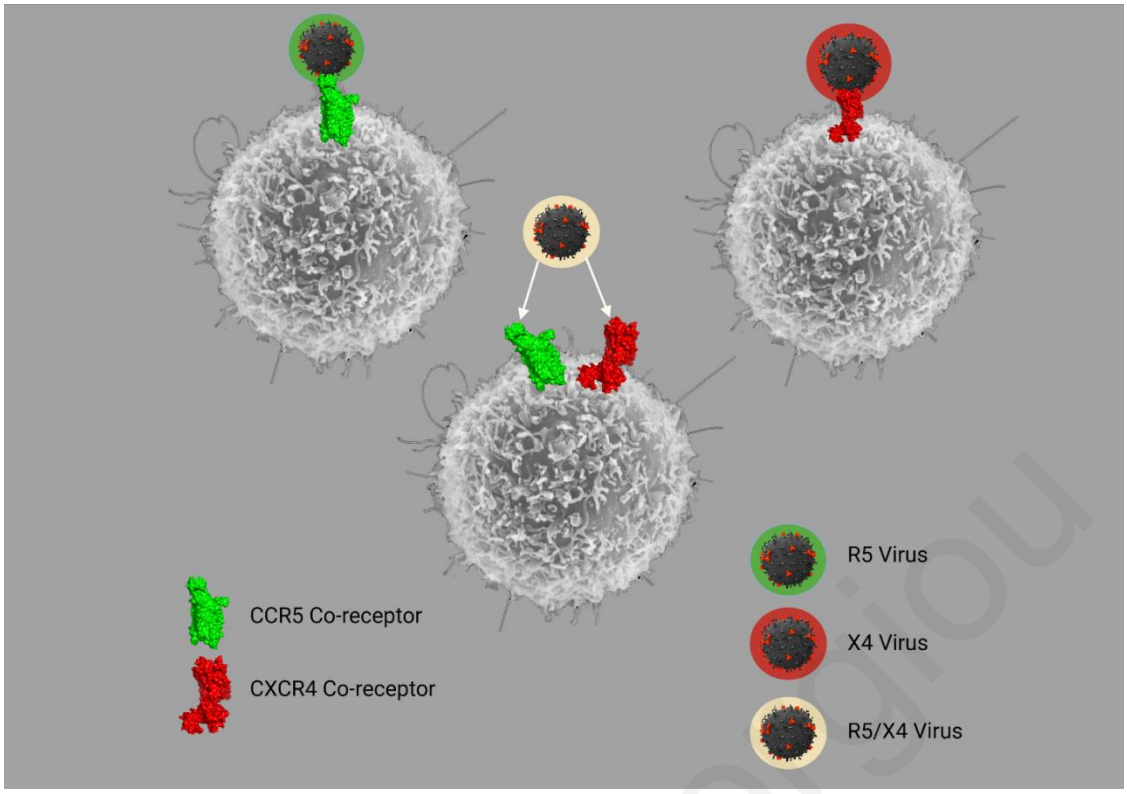


Figure 4: *Schematic representation of the HIV-1 Tropism.* The green color represents the R5 strains of the virus that can infect cells with the CCR5 co-receptor, and the red color represents the X4 strains of the virus that can infect cells with the CXCR4 co-receptor. The skin color represents the R5/X4 strains of the virus that can infect cells with either CCR5, CXCR4 or both co-receptors. CCR5 PDB entry: 4mbs; CXCR4 PDB entry: 3OE0; PDB=Protein Database.

1.4 HIV-1 Transmission

HIV-1 transmission is characterized as mainly horizontal and can occur only through specific body fluids. These fluids are blood, semen, pre-semen, anal fluids, vaginal fluids, and breast milk. Due to the means of transmission, for the transmission of the virus to be possible, these fluids must come into contact with genital and serum mucosa, damage tissue, or by inject the virus direct into the bloodstream with a syringe or needle. Because of the means that the virus uses for transmission, the most common routes of infection of the HIV-1 are unprotected sex and the exchange of needles mainly between heroin users. Depending on the temperature and other factors, the virus can live on a used needle for up to 42 days. As heroin users often exchange the same needle, this way of transmission is one of the main route of transmission after the unprotected sex. In addition, the virus can be transmitted vertically from the mother to the child during childbirth or breastfeeding (Centers for Disease Control and Prevention 2022, Kreiss 1997, Rom, Markowitz 2007, Fox, Fidler 2010).

1.5 Life Cycle of HIV-1

The HIV-1 reproductive cycle is complex and can be described in several stages. The initial step is the binding to the surface of the host cell followed by its penetration and release of the capsid into the cytoplasm. This step is accomplished by direct fusion of the viral envelope with the host membrane. Next is the reproduction, which occurs after the viral capsule disintegrates and releases its genome and key enzymes (RT, IN and pol genes that encode enzymes). At this stage, the RNA genome is reverse-trascribed with reverse transcriptase into double-stranded DNA, and then it is integrated into the genome of the host with the help of integrase. After the transfer and integration of the viral double-stranded DNA into the host DNA, the transcription and translation of the viral genes take place by using the host mechanisms. Following the transcription and translation of the viral genes, is the assembly, in which the viral capsule is assembled into an immature form. Subsequently, the viral particle acquires its envelope, from the host membrane. The final step is maturation and release. At the stage of maturation, which occurs during release, the viral gag and pol proteins are broken down by retroviral protease, thus forming the mature and infectious form of the virus (Saxena, Chitti 2016, Kirchhoff 2013) (Figure 5).

To achieve attachment and entry, the virus uses the envelope proteins gp120 for the CD4 receptor recognition and binding, and gp41 protein for the binding to CCR5/CXCR4 co-receptors (detailed description in annex 1.3). Overall result of the receptor (CD4) and co-receptor (CCR5, CXCR4) recognition from the HIV-1 envelope proteins (gp120, gp41) is the dramatic rearrangement of gp41 that leads to the fusion of viral envelope membrane with the host cell membrane (Chan, Kim 1998, Wilen, Tilton et al. 2012). In such a way the reverse transcriptase complex (MA,VPR,RT and IN) is released into the cytoplasm.

Next is the DNA synthesis. For the implementation of this stage, the use of RT is necessary. Reverse transcriptase is an enzyme, which act as a DNA polymerase and can use both DNA and RNA as a template, and also has the ability to act as an RNase H (Hu, Hughes 2012). Reverse transcriptase uses the single-stranded RNA of the virus as a template, and after a series of complex steps, transforms the RNA into double-stranded DNA (Götte, Li et al. 1999, Hu, Hughes 2012). This stage has major effects on the current HIV-1 genetic diversity, which is further explained in annexes 1.7.1 and 1.7.2.

Then, the integration of viral DNA into the host DNA is performed. This process involves the formation of a pre-integration complex, that lead to the integration of the viral DNA into the chromosomal DNA. For the implementation of this step, integrase protein is required (Piller, Caly et al. 2003). Upon completion of this step, the cell remains infected for the rest of its life. Additionally, as part of the chromosomal DNA, proviral DNA is replicated by host cell proliferation, thus the infection can be achieved either by infecting new cells or by proliferation of the already infected cells (Kirchhoff 2013)

Subsequently, the transcription of the virus DNA is executed through the cell transcription engine (RNA pol II, NF-kB, NFAT transcription factors). As the cell cannot differentiate the viral DNA, it expresses it as one of its own normal genes. For this procedure except of the cell transcription engine, the virus proteins are also needed. For instance, Tat protein binds to the proviral DNA and with the cellular transcription factors (pTEFb) enhances the speed of transcription. The transcripts are then subjected either to complete (2 kb RNA encoding Tat, Rev, and Nef) or partial (4 kb RNA encoding Vif, Vpr, Vpu, and env) alternative splicing. If the transcript does not undergo alternative splicing, then the product of transcription is the viral genome or the premature genes of gag and gag-pol regions. The transfer of transcripts from the nucleus to the cytoplasm for translation is done with the help of Rev protein (Kirchhoff 2013). It should be noted that mRNAs are produced according to time necessity of each protein, depending on the function it performs and the time at which it is needed (Wu, Y., Marsh 2003). It is also important to mention that the change in expression of HIV-1 genes occurs due to different sites of integration in genome and consequently the activity of regulatory proteins in close chromosomal positions. The negative regulation of HIV-1 transcription is particularly significant because of its association with the latent status of HIV-1 in cells of HIV-1-infected patients (Lewinski, Bisgrove et al. 2005).

Finally, the assembly, maturation, and the budding are executed. The assembly is achieved by polymerization between the gag and gag-pol precursor molecules, and through interactions (N-terminally myristoylation) with the matrix, these molecules are concentrated in lipid platforms inside the cell membrane (Li, H., Dou et al. 2007, Ganser-Pornillos, Yeager et al. 2008). Additionally, the glycoproteins of the envelope are recruited on the lipid platforms through the interaction with matrix proteins. Then, the two copies of viral RNA genome are recruited to the complex through interactions with capsid proteins (NCs). Finally, the recruitment of Vif protein

to the complex occurs, which plays an important role as it acts as a competitor for cellular enzymes (APOBEC3F and APOBEC3G), and thus, exclude the possibility of their assembly in the virus. The completion of recruitment of all agents to the cell membrane induce the membrane curvature and subsequently the formation of a spherical virus particle that is encapsulated within the host membrane. During this process the dimers gag-pol cause their own proteolysis, thus producing various enzymes including protease. The protease then cleaves the precursors of gag and gag-pol creating the final mature components of the virus, thus making it contagious (Cadima-Couto, Goncalves 2010, Sundquist, Krausslich 2012)

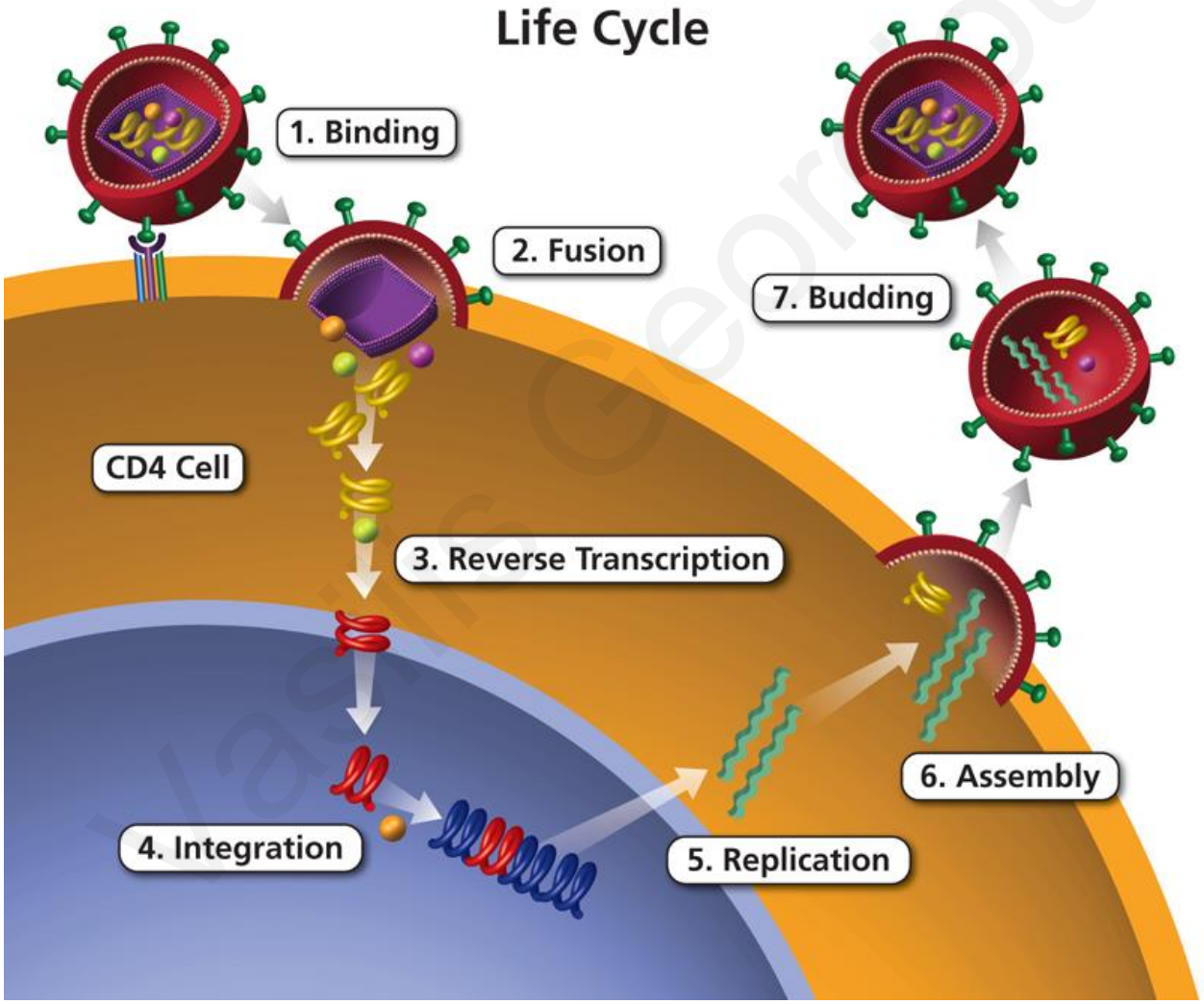


Figure 5*: Life cycle of HIV-1.* The schematic representation illustrates the life cycle of the HIV-1 virus divided into 7 crucial steps (National Institute of Health 2011).

1.6 Antiretroviral Therapy (ART)

The first approach to the development of antiretroviral therapy began in summer of 1985, where in vitro experiments described the inhibition ability of AZT (zidovudine, 3 '-azido-2', 3 '-didoxythymidine), which subsequently was the first approved antiretroviral drug for clinical use against HIV-1 (Mitsuya, Weinhold et al. 1985). Soon in March 1986, other variants of AZT were described and in particular the DDI and DDC drugs (Mitsuya, Broder 1986). The addition of other drugs in the same category as AZT and the discovery of new ART classes led to the current era of combined ART, also known as Highly Active Antiretroviral Therapy (HAART), that significantly increased life expectancy. These drugs are categorized in six distinct categories based on their molecular mechanism and resistance profile. These classes are (1) nucleoside reverse transcriptase inhibitors (NRTIs), (2) non-nucleoside reverse transcriptase inhibitors (NNRTIs), (3) integrase inhibitors (INSTIs), (4) protease inhibitors (PIs), (5) entry inhibitors and (6) co-receptor antagonists (Arts, Hazuda 2012, Santoro, Perno 2013) (Figure 6). Their combined action results in the dramatic reduction of viral proliferation that lead to the decrease of HIV-1 virus in the plasma to a point that is undetectable, which allows the significant recovery of the immune system (Autran, Carcelain et al. 1997, Komanduri, Viswanathan et al. 1998, Lederman, Connick et al. 1998). Additionally, the expansion of global access to ART has allowed 28.2 million people living with HIV to have access to treatment (statistics corresponding until the end of June 2021) (Unaids, 2022). Furthermore, in addition to extending accessibility, new strategies are being adopted, which are aimed at reducing the transmission of the virus. For example, one of the rapidly evolving strategies is antiretroviral pre-exposure prophylaxis (PrEP). PrEP, according to the World Health Organization (WHO), (https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/prevention/pre-exposure-prophylaxis) is a method of prevention for individuals that are not infected with HIV-1, but belong to a high-risk group, where they receive daily antiretroviral therapy in order to reduce the risk of future HIV infection. Reaching today, the target of WHO is 90:90:90. It stands for 90% of the people who live with HIV-1 to know it, 90% of people diagnosed with HIV-1 to receive ART, and 90% of people receiving ART to suppress the virus. The ultimate goal of WHO is to maintain high levels of viral suppression in the population,

thus reducing the risk of transmission resulting in control of the epidemic of the virus. Based on UNAIDS HIV and AIDS fact sheet (https://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf), the goal of the WHO is almost complete and specifically in 2020, 84% of people living with HIV knew their HIV status, among those people who knew their status, 87% were accessing treatment and among the people accessing treatment, 90% achieved viral suppression.

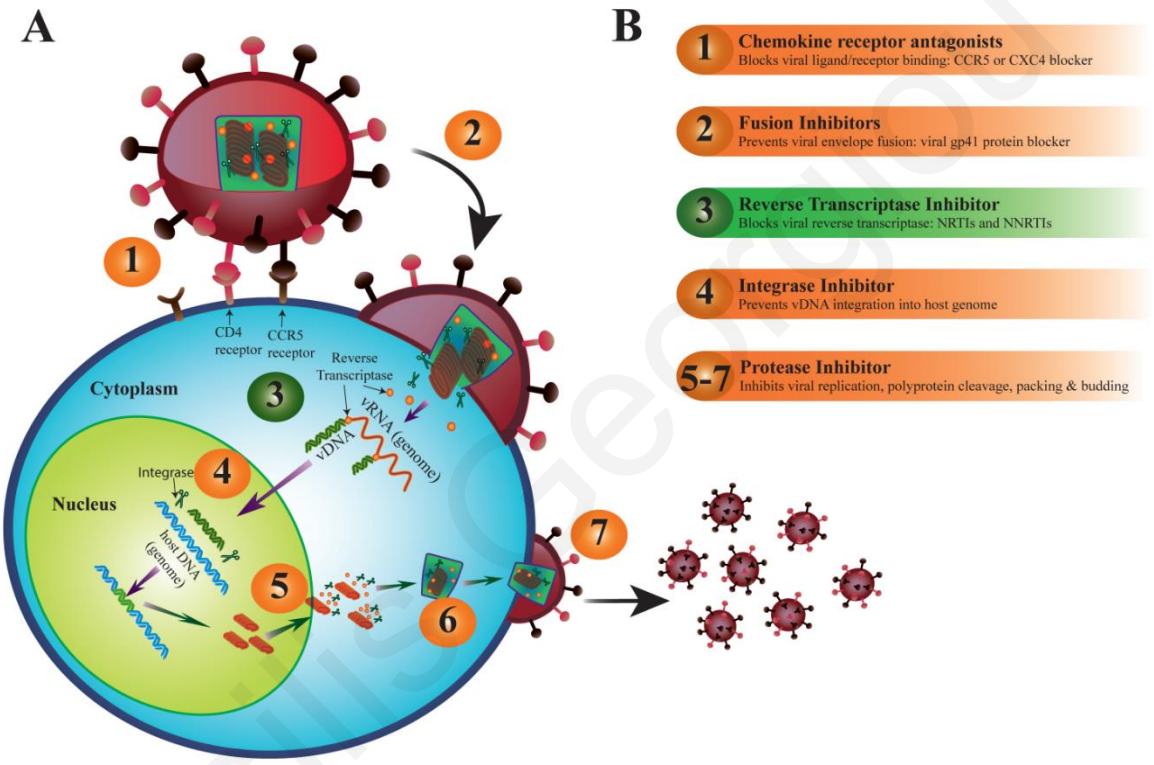


Figure 6: *Schematic representation of the antiretroviral treatment* categories, *based on the life cycle of the HIV-1.* (A) Illustrates the life cycle of the HIV-1 and (B) shows the categories of the antiretroviral drugs based on their role on the life cycle of HIV-1. (Holec, Mandal et al. 2017).

1.7 Genetic Hypervariability of HIV-1

HIV-1 was described in the early 90s and yet it is one of the most challenging viruses ever described in terms of the fight for its extinction. Throughout the years, it became clear that the main reason that makes HIV-1 so successful as a virus, it's the genetic hypervariability. The broad genetic diversity of the virus is caused by various factors. The most important factors are the ability of reverse transcriptase to switch templates (Hu, Temin 1990, Smyth, Davenport et al. 2012)

(annex 1.7.1), the high error rate that this enzyme presents (Mansky, L. M., Temin 1995, MANSKY, LOUIS M. 1996) (annex 1.7.2), as well as the high viral turnover (Ho, Neumann et al. 1995). As a result of the aforementioned factors, over the years, the HIV-1 evolved into several subtypes and numerus recombinant forms, which has been transmitted all over the world (Hemelaar, Elangovan et al. 2019, Hemelaar, Elangovan et al. 2020) (annex 1.7.3 and 1.7.4). Because of the HIV-1 genetic hypervariability, a repertory of implication in multiple aspects has been described, which is analyzed in detail in the annex 1.8.

1.7.1 Recombination of HIV-1

It has been clear through the scientific literature that a single RNA strand is sufficient for the replication of the HIV-1 as it contains all the genetic information that is required. Nevertheless, most HIV-1 virions contain two copies of the viral RNA (Chen, Nikolaitchik et al. 2009). The co-packaging of two RNA strands into the virion increases the possibility of recombination and therefore drive the genetic hypervariability that characterizes the HIV-1(Temin 1991). The genetic recombination enables the virus to escape the immune responses and the antiretroviral treatment, as well as the lethal mutated genomes of the virus, which may be generated through the extremely high mutation rate (Bull, Sanjuan et al. 2007, Quan, Liang et al. 2009). For an HIV-1 recombination event to occur, two or more HIV-1 virions must infect the same cell. Then, the two distinct HIV-1 RNA genomes are reversed transcribed and integrated into the host genome. Following that, the proviral DNA genomes of both are transcribed through the host mechanism and the expressed RNA strands of both HIV-1 virions are packaged into a budding virion. After that, the new virion that contain the heterodimeric RNA genome must infect another cell and during the procedure of reverse transcription the HIV-1, RT must switch templates between the two heterodimeric RNA genomes and in such a way to produce a new recombinant genome that will later be integrated into the host genome and expressed the new recombinant RNA genome (Yaseen, Abuharfeil et al. 2017) (Figure 7).

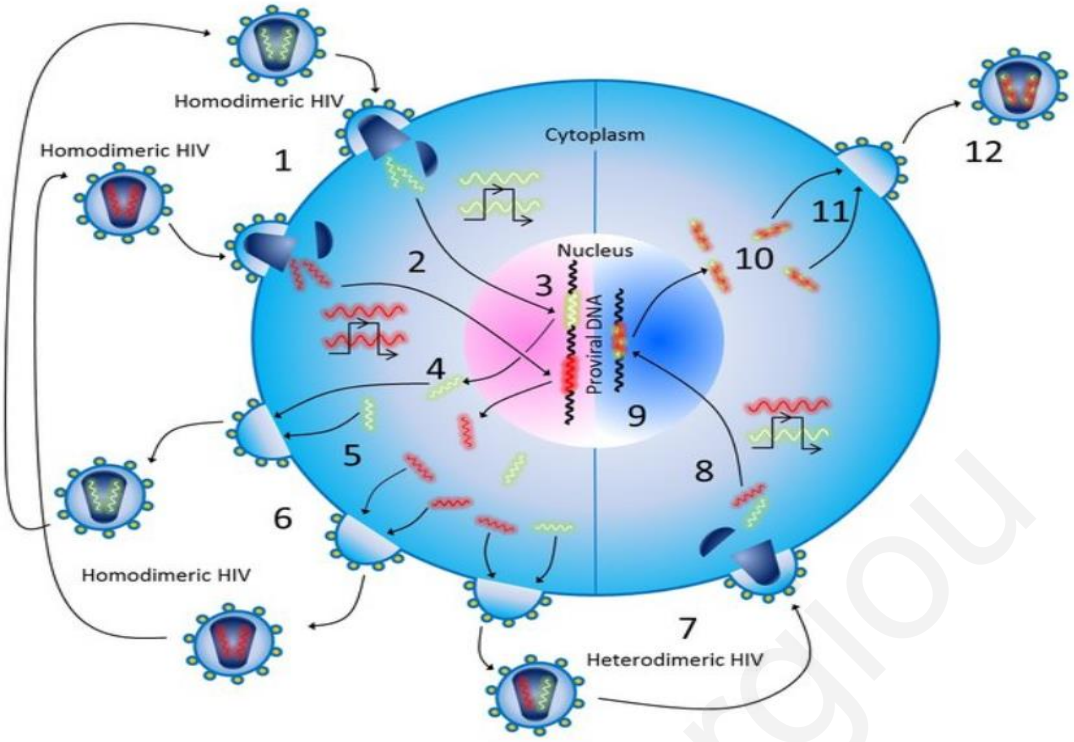


Figure 7: *Recombination of HIV-1*. The schematic representation illustrates two distinct cells by using two different colors of the nucleus as well as with the vertical line that divide the cell into two segments. The left part of the cell (pink nucleus) is the initial cell that get coinfected with two homodimeric HIV virions represented with the green and red colors. The right part of the cell (blue nucleus) represents the second cell that get infected with the heterodimeric HIV virion that was created as a result of the initial coinfection with the two homodimeric HIV virions. In detail for an HIV-1 recombination event to occur (1) two distinct HIV-1 homomeric virions must infect the same cell (2) and reverse transcribed their genomes. Following that (3) the reversed transcript DNA proviral genomes must be integrated into the host genome and then (4) to be expressed through the host cell RNA transcription machinery. After that (5) the new RNA strands will co-packaged into either (6) the initial homodimeric HIV-1 virions or (7) a new heterodimeric HIV-1 virion or both. Next (8) the new heterodimeric virion will infect another target cell and through a switching template event during the reverse transcription processed a new HIV-1 recombinant genome will be produced. Then through the steps 9-12 a new recombinant HIV-1 particles will appear (Yaseen, Abuharfeil et al. 2017).

Focusing on the step 8 of the Figure 7 (template switching) the general mechanism is the following. During reverse transcription, the RT activity as RNase H, degrades the genomic RNA that has already been copied, thus forming the RNA donor, creating in such a way an extended area of a negative single-stranded DNA (-ssDNA). In the presence of appropriate stimulation, the transfer of the -ssDNA 3'OH prime to the RNA recipient may be induced, resulting in template switching (Figure 8) (Smyth, Davenport et al. 2012). Suitable stimuli can be present in areas that RT pauses, or the secondary structures of the genomic RNA. (DeStefano, Bambara et al. 1994, Wu, W., Blumberg et al. 1995, Roda, Balakrishnan et al. 2002, Derebail, DeStefano 2004). The secondary structures of genomic RNA facilitate the transfer of -ssDNA to the RNA receptor using a mechanism similar to Holliday junctions mechanism, which normally occur during DNA recombination (Moumen, Polomack et al. 2003, Galetto, Giacomoni et al. 2006).

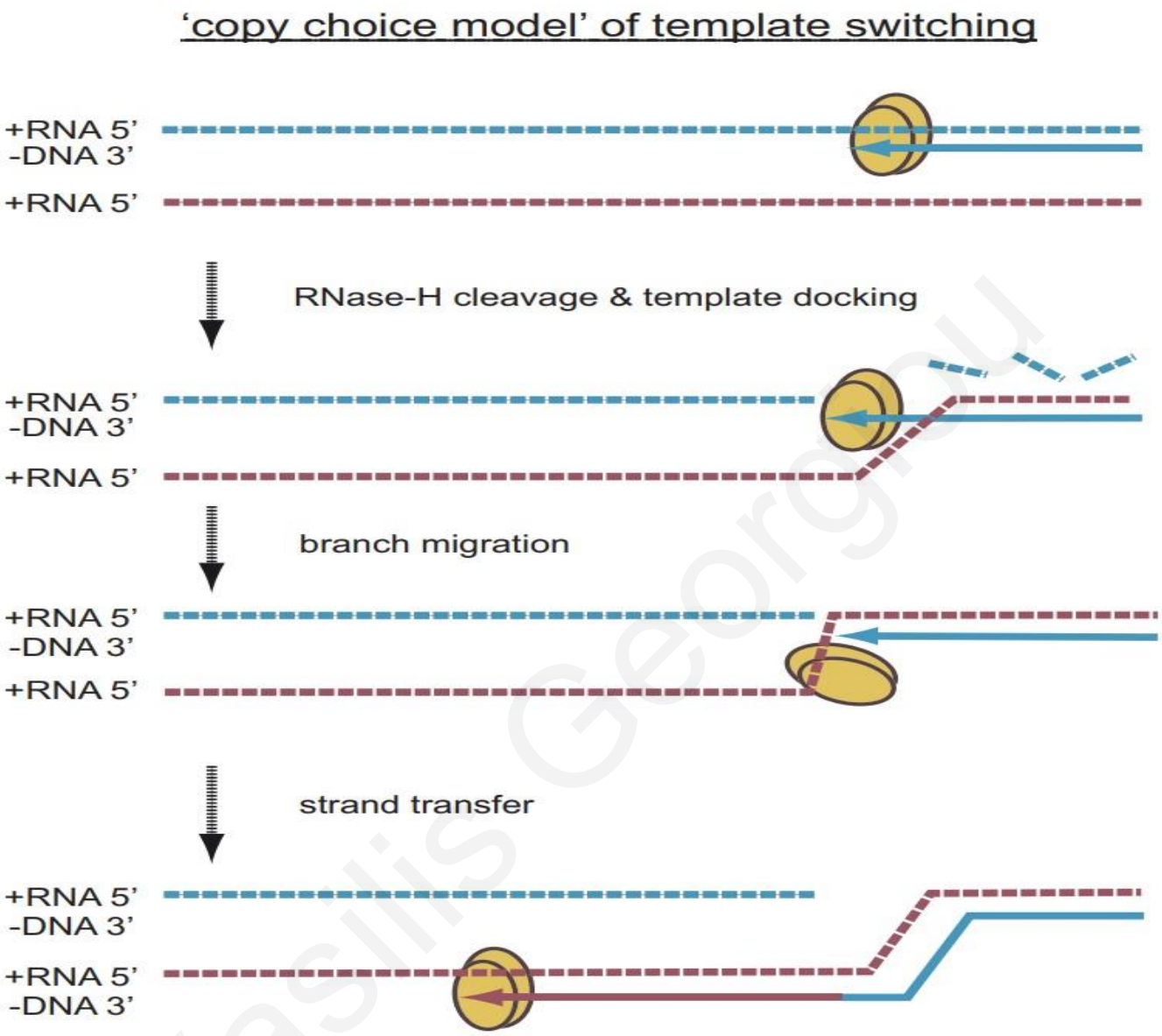


Figure 8: *Template switching model.* Template switching is occurred during the synthesis of the -ssDNA through the activity of the RT as an RNase H. The degradation of the RNA from the RNA/DNA hybrid exposes the monoclonal regions of -ssDNA and as a result can base pair with the complementary RNA genome (indicated with the red color). For the completion of the template switching model a branch migration and a strand transfer are follow (Smyth, Davenport et al. 2012).

1.7.2 Reverse Transcriptase Error Rate

Reverse transcriptase is an enzyme that presents a very high error rate due to the lack of proofreading. The sequences of HIV-1 between different patients but also within the same patient varies significantly. Despite that most patients are infected with a single virus, the broad genetic variation that is later detected is remarkable, thus, is clear that the variation that was detected and as such the evolution of the virus occurred after the infection (Keele et al. 2008). This is a result of the high turnover and the error rate of reverse transcriptase that is relatively high causing numerus mutations during the life cycle of the virus (Coffin 1995). Specifically, research conducted by Robert et al. has shown that using in vitro experiments the reverse transcriptase error rate is $5 * 10^{-4}$ mutation per base when RT uses DNA as a template (Roberts, Bebenek et al. 1988). However, it was later shown that the error rate resulting from the reverse transcriptase activity was much lower. Specifically in a non-viral reference sequence, after a cycle of reproduction in T-cell lines, the error rate was $3 * 10^{-5}$ errors per base, which is less than 5% of the in vitro experiments described above (Mansky, L. M., Temin 1995, MANSKY, LOUIS M. 1996). Regardless, the error rate remains particularly high compared to DNA polymerase, which has the ability of proofreading and shows an error rate of 10^{-10} mutations per base (Preston, Albertson et al. 2010). Nevertheless, there is limited data with respect to the proportion of the mutations resulting from the HIV-1 reverse transcriptase or the host DNA dependent RNA polymerase II (RNA pol II), which is another enzyme lacking proofreading ability that has a crucial role in the life cycle of the virus (Hu, Hughes 2012).

1.7.3 Subtypes of HIV-1 and Geographic Distribution

Since the zoonotic transmission that has been described in the annex 1.1, HIV-1 has been evolving for more than three decades. As a result, the HIV-1 M group has been divided into ten distinct pure subtypes (A, B, C, D, F, G, H, J, K and L). Subtype L has been recently described as part of the work of Yamaguchi *et al.*, using next-generation sequencing, in March 2020 (Yamaguchi, Vallari et al. 2020). Since the beginning of the pandemic of HIV-1, different subtypes dominate different regions of the world at different time points. From 1990 to 2015, subtype B was responsible for 12.1% of infections, followed by subtype A 10.3%, CRF02_AG 7.7%, CRF01_AE 5.3%, subtype G 4.6%, and subtype D 2.7%. The combination of subtypes F, H, J, and K accounted for 0.9% of infections. The rest of the CRFs reported for 3.7% leading to the total proportion of all CRFs to 16.7%. The URFs constituted 6.1%, resulting in the total proportion of 22.8% for all the recombinants globally. The global distribution of HIV-1 subtypes and recombinants has been unstable over time and changes appear in countries, regions, and proportions. At a global level for the period from 2005 to 2015, the prevalence of subtype B increased, subtypes A and D were stable, and subtypes C, G and CRF02_AG decreased. The prevalence of CRF01_AE, other CRFs, and URFs continued to increase, leading to a consistent increase in the global proportion of recombinants over time (Hemelaar, Elangovan et al. 2019). All the subtypes are present in sub-Saharan Africa and mainly in central-western Africa, where the highest percentage of genetic diversity is observed (Buonaguro, Tornesello et al. 2007). Specifically, in central Africa, all HIV-1 subtypes and many CRFs and URFs were found, between 1990 and 2015 (Hemelaar, Elangovan et al. 2019). In Asia (India), subtype C was the predominant subtype, that contributed for at least 89% of the infections from 1990 to 2015. The main subtype in western and central Europe, North America, Caribbean, Latin America, and Oceania with over 75% of the infections from 2010 to 2015 was subtype B. In eastern Europe and central Asia, subtype A accounted for more than 50% of the HIV-1 infections (Hemelaar, Elangovan et al. 2019) (Figure 9).

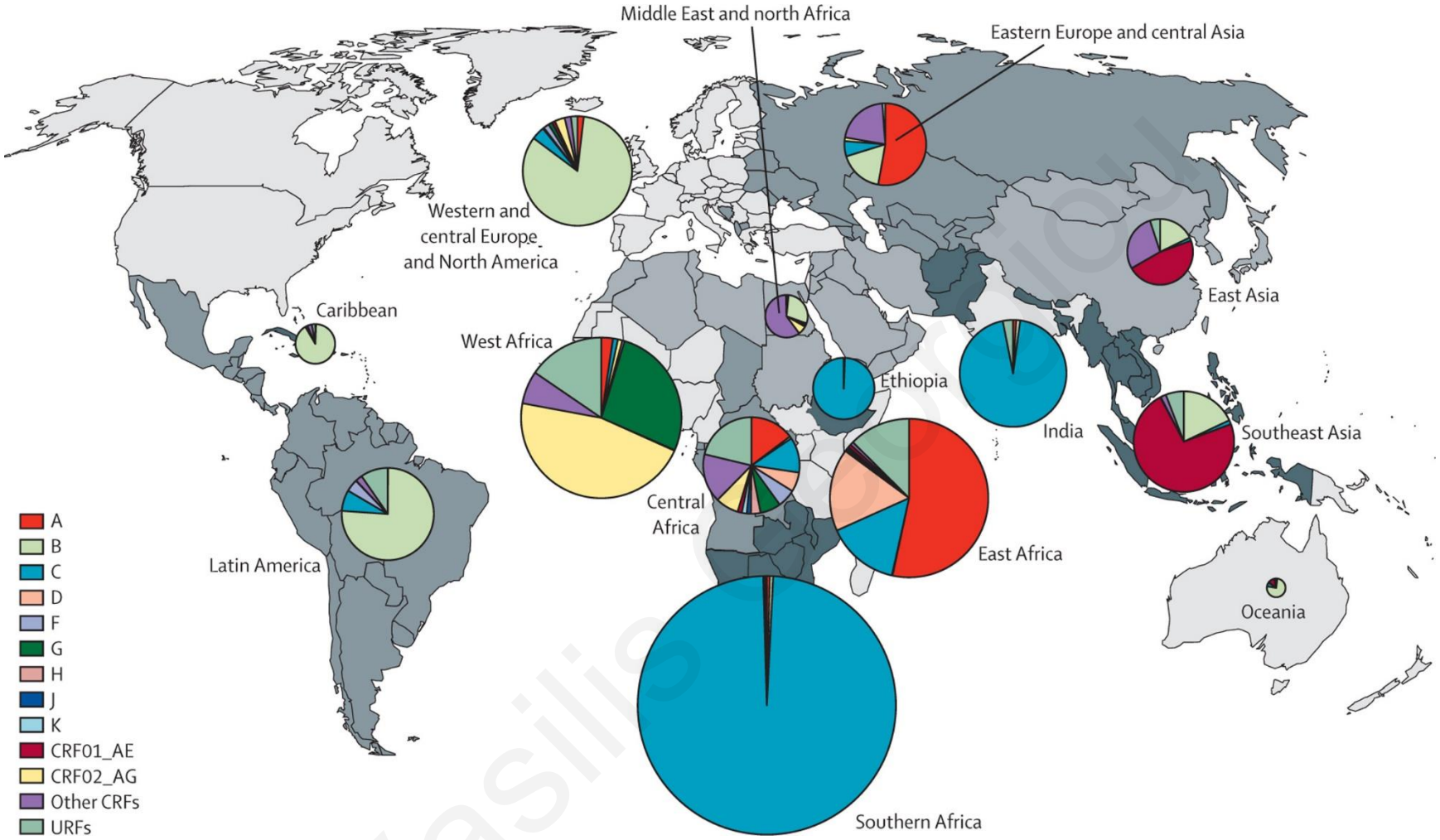


Figure 9: *Global distribution of HIV-1 subtypes CRFs and URFs from 2010 until 2015*. The countries were grouped into 14 regions indicated with the different gray colors of the map. The size of the pie-chart illustrates the proportion of people that live with HIV-1 in each region and the colors of the pie chart indicates the subtype proportion for each region. CRF=circulating recombinant form. URFs=unique recombinant forms (Hemelaar, Elangovan et al. 2019).

1.7.4 Recombination Forms of HIV-1 and Geographic Distribution

The existence of numerus subtypes in the same geographic areas, could result in simultaneous infection of individuals with viruses belonging to different subtypes (Yaseen, Abuharfeil et al. 2017). Thus, due to recombination events (detail description in annex 1.7.1), novel HIV-1 recombinant strains arise over time. The region in the recombinant genome where a transition from a certain subtype to another subtype is detected, is defined as a "breakpoint" (Compendium 2018). The HIV-1 recombinants are categorized into two distinct groups, the CRFs, and the URFs (Robertson, Anderson et al. 2000). A URF is an inter-subtype recombinant genome that was found only in one dually infected (or multiply infected) individual patient (Robertson, Anderson et al. 2000). A CRF is an inter-subtype recombinant virus that was transmitted to many people, and it became one of the circulating strains in the HIV epidemic (Robertson, Anderson et al. 2000). For the identification of a CRF, at list three identical recombinant HIV-1 genomes that have infected three or more people, who are not epidemiologically related, must be detected (Compendium 2018). The CRFs are labeled with numbers in the order in which they were first adequately described and currently in the Los Alamos HIV database 118 CRFs have already been described (Compendium 2018). The prevalence of different CRFs among different geographic regions around the world vary in the same manner as the HIV-1 subtypes. A recent study found that the proportion of recombinants increased over time, both globally and in specific geographic regions, reaching 22.8% of global HIV-1 infections from 2010 to 2015 (Hemelaar, Elangovan et al. 2020). Precisely the proportion of the CRFs and the number of newly identified CRFs increased over time to 16.7% and 57 CRFs, respectively, from 2010 to 2015. Globally, the most prevalent recombinant was CRF02_AG, accounting for 33.9% of all recombinant infections from 2010 to 2015. URFs accounted for 26.7%, CRF01_AE for 23.0%, and other CRFs for 16.4% of all recombinant infections from 2010 to 2015 (Hemelaar, Elangovan et al. 2020). The regional distribution and the quantity of recombinants form 2010 until 2015 is indicated in the Figure 10 below.

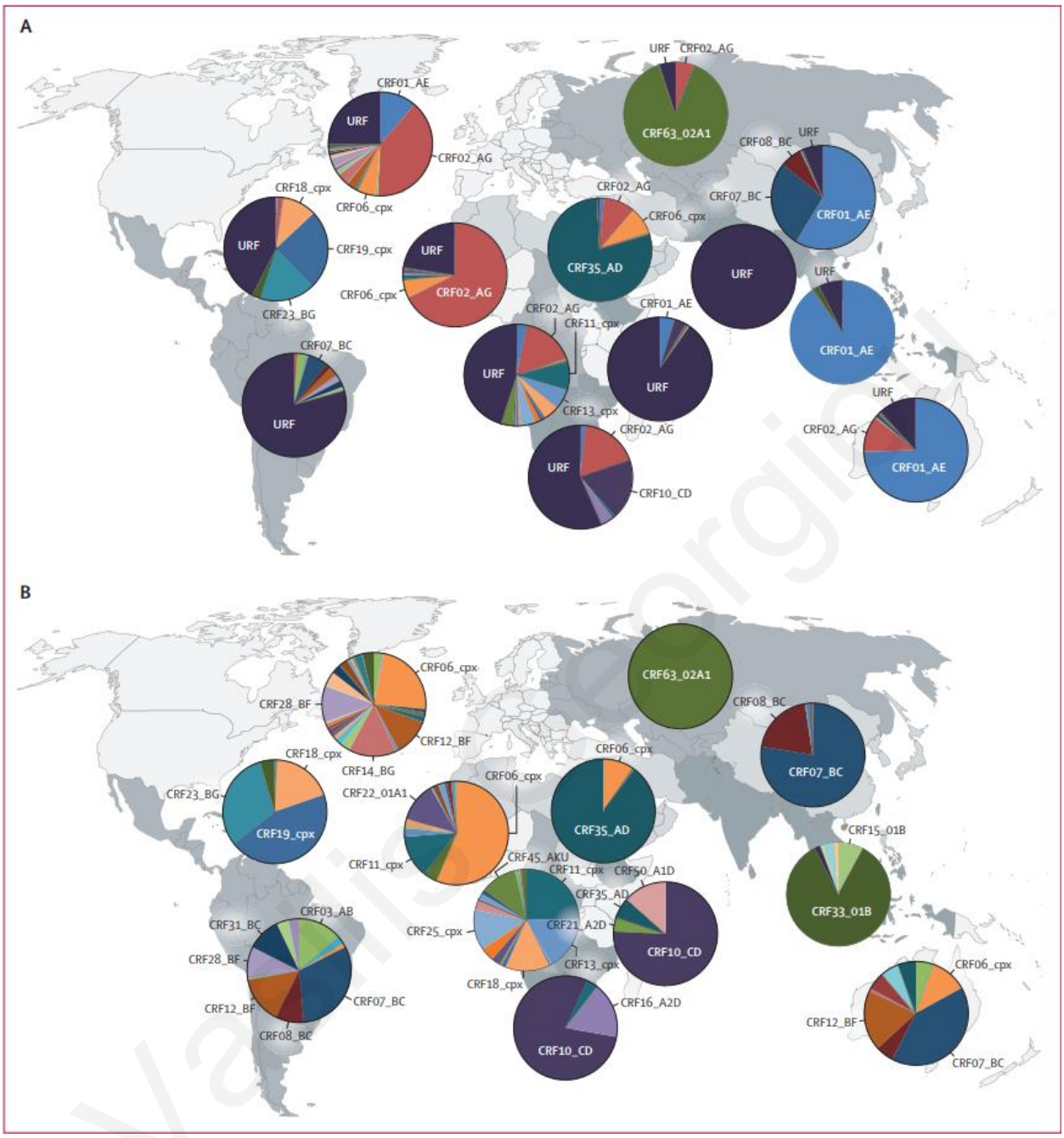


Figure 10: *Global distribution of the HIV-1 recombinants* from *2010* to *2015*. The counties are grouped into 14 regions indicated with the different gray color coding on the world map. The colors of the pie charts indicates the quantities of recombinants for each region. (A) The scheme above illustrates the global distribution of all the HIV-1 recombinants (URFs and CRFs). (B) The scheme below shows the global distribution of all the HIV-1 recombinants (URFs and CRFs), excluding the predominant CRF01_AE and CRF02_AG. CRF=circulating recombinant form. URF=unique recombinant form (Hemelaar, Elangovan et al. 2020).

1.8 Implications Derived from the HIV-1 Genetic Diversity

There is no doubt that the HIV genetic diversity has several implications on various aspects including transmission routes, disease progression, drug resistance that effect response to HAART, viral load measurement, diagnostics, pathogenesis, immune response and vaccine development. In this section there is an overview of the current knowledge about the impact of the HIV variability on the antiretroviral therapy (annex 1.8.1), disease progression (annex 1.8.2) and vaccine development (annex 1.8.3).

1.8.1 Resistance to Antiretroviral Therapy Among Different HIV-1 Subtypes.

Over the years the HIV-1 has acquired a very broad genetic diversity. This diversity has been implicated in the efficiency of the antiretroviral therapy among different subtypes and CRFs. The scientific community manage to reveal different sensitivity levels and several drug resistance mutations that are distinct for some subtypes and recombinants. The initial research that aimed to reveal subtype specific drug resistance mutations was conducted in the late 90s. In particular, a study that was performed in 1998 by Palmer *et al.* showed that the clade D function with diminished drug sensitivity and has rapid growth kinetics, whereas subtypes A, B, C and E demonstrated comparable sensitivity (PALMER, ALAEUS et al. 1998). Another study that was conducted in the same period of time by Apetrei *et al.* discovered that clade F shows some measure of resistance to the non-commercialized NNRTI, the TIBO compound (Apetrei, Descamps et al. 1998). Later, in 2002, Loemba *et al.* through their results indicate that clade C isolates show inherent resistance against NNRTIs due to the presence of a single G190A mutation (Loemba, Brenner et al. 2002). Additionally, a more recent study that was conducted in 2010 in Nigeria also indicate different subtype specific drug resistance mutations of the RT and PR among the subtypes A, G, CRF02_AG and CRF06_cpx (Chaplin, Eisen et al. 2011). Reaching today's data, as already has been described in annex 1.7.3 and 1.7.4, the recombinants strains are increasing globally, leading in such a way, to a more complex surveillance in terms of the subtype specific antiretroviral drug resistance mutations. The overall conclusion is that there are differences in the sensitivity to antiretroviral therapy that are associated with the HIV-1 subtypes. This has obvious implications on the antiretroviral treatment against HIV-1.

1.8.2 Consequences of the HIV-1 Genetic Diversity on Disease Progression

Throughout the years, from the beginning of the HIV-1 pandemic, until today, numerus studies examined whether there is an association between HIV-1 subtypes and disease progression. The earliest studies that were conducted between 1997-1999 concluded no correlation between viral subtypes and disease progression (Alaeus, Lidman et al. 1999, Amornkul, Tansuphasawadikul et al. 1999, Galai, Kalinkovich 1997). However, recent studies did find significant differences (Kiwanuka, Robb et al. 2010, Baeten, Chohan et al. 2007, Kaleebu, French et al. 2002, Ssemwanga, Nsubuga et al. 2013, Vasan, Renjifo et al. 2006, Venner, Nankya et al. 2016, Amornkul, Karita et al. 2013). Direct comparison between subtypes is challenging and hence these contradictory conclusions could be due to differences in cohort size, disease stage of the patients, host and environmental factors, subtype identification methods or other factors. Taking those factors into consideration, the studies that examine pathogenesis should be conducted in populations where co-circulation of different subtypes exist, as this could potentially eliminate some of these confounding variables. One example is a prospective study involving Senegalese female sex workers that revealed that women who were infected with non-subtype A virus had a greater risk of developing AIDS rather than subtype-A-infected women in the same cohort (Kanki, Hamel et al. 1999). Similarly, Amornkul *et al.* assessed disease progression in a cohort of individuals infected with subtypes A, C, and D, and found comparable results, where individuals infected with non-subtype A had faster CD4 decline compared to individuals infected with subtype A (Amornkul, Karita et al. 2013). Another study that was based in a seroincident cohort in Uganda revealed that individuals infected with subtype D had a faster CD4 decline when compared to subtype A-infected individuals (Kiwanuka, Robb et al. 2010). Likewise, Baeten *et al.* also observed an association between CD4 decline and the HIV-1 subtypes, in a cohort of commercial sex workers in Kenyan. The results of this study revealed that there is a higher risk of disease progression leading to death in individuals infected with subtype D compared to subtype A, and this risk persisted after viral suppression (Baeten, Chohan et al. 2007). Equally, Ssemwanga *et al.* found that subtype D-infected individuals had faster CD4 decline when compared to subtype A-infected individuals in rural Uganda (Ssemwanga, Nsubuga et al. 2013). Furthermore, subtype D infection was associated with the greatest CD4 cell decline in a cohort of women in Sub-Saharan Africa, who were either infected with subtype A, C, or D (Venner, Nankya et al. 2016). Wymant

et al. report an extremely virulent subtype of HIV that has been circulating in the Netherlands for more than a few years, by examining data of well-characterized European cohorts. This distinctive subtype B variant of HIV-1 was reported to have double the rate of expected CD4+ cell count declines, and as a consequence, the individuals that were infected with this strain were vulnerable to developing AIDS within 2 to 3 years (Wymant, Bezemer et al. 2022). The general conclusion drawn from the literature is the evidence of differences in disease progression associated with viral subtype. This has important implications for the treatment of HIV-1 infection, and therefore the life expectancy of the HIV-1-infceted patients.

1.8.3 HIV-1 Genetic Variety and Vaccine Development

Historically the vaccines have been one of the most important tools for ending viral epidemics with several great examples over time (Wadman, You 2017). In the case of the HIV-1 virus, the vaccine development has been proven to be a huge challenge for the scientific community for several reasons. The major unprecedented challenges for vaccine development are the exceptional diversity of HIV-1, the ability of the virus to avoid adaptive immune responses, the lack of ability to stimulate broadly reactive antibody responses and the early establishment of latent viral reservoirs (Barouch 2008). In this section the focus will be in the aspect of the HIV-1 genetic variety implications for the vaccine development. There is no doubt that the global HIV-1 hypervariability account for one of the major barriers to vaccine development against HIV-1 as a globally effective HIV-1 vaccine has to have the ability to protect against divergent HIV-1 subtypes and recombinants (Barouch, Korber 2010). Thus, the up-to-date and accurate knowledge of the global distribution of HIV-1 subtypes and recombinant is essential especially for subtype specific vaccines because vaccine immunogen sequences need to be as similar as possible, to the viral sequences circulating in the target population (Gaschen, Taylor et al. 2002). Currently, there are several vaccine trials attempts, with the most clinically advanced, that demonstrate sufficient safety and immunogenicity, to be reported in phase 1 and 2a in humans trials (Bekker, Moodie et al. 2018, Barouch, Tomaka et al. 2018). One of those vaccines that entered the efficacy trial (HVTN705; NCT03060629), which take place in southern Africa, is a polyvalent mosaic vaccine designed to cover all global HIV-1 group M viruses. On the other hand the second vaccine that entered the efficacy trials (HVTN702; NCT02968849), which take place in South Africa, investigates a vaccine that was developed based on the vaccine used in the successful RV144 trial

in Thailand (Rerks-Ngarm, Pitisuttithum et al. 2009). This vaccine trial has undergone a curtial adjustment, which was the replacement of the subtype B/CRF01_AE immunogens with subtype C isolate immunogens, that match the predominant HIV-1 subtype in South Africa. If this adjustment proved to be effective against subtype C infection, the next step would be to check if the vaccine has the ability of cross-protection against various HIV-1 subtypes and recombinants by testing the vaccine in regions of the world that are characterized by several subtypes and recombinants such as east Africa. However, whether this vaccine is sufficient to protect against these diverse subtypes and recombinants or if it will be a subtype-specific vaccine is still unknown (Barouch, Korber 2010, Barouch, Tomaka et al. 2018). This vaccine leans towards influenza like vaccine that might need to be changed periodically according to the global distribution of HIV-1 subtypes and recombinants (Korber, Gaschen et al. 2001). The overall conclusion is that the identification of novel HIV-1 subtypes and recombinants as well the surveillance of the global distribution of the HIV-1 is essential for the vaccine development. Additionally, the accurate regional and global prevalence of HIV-1 subtypes and recombinants is also vital to be known for the estimation of the global market need for therapeutic and prophylactic HIV-1 vaccines, allowing the prioritization of the subtypes with the greatest potential global benefit (Marzetta, Lee et al. 2010)

1.9 Aim of the Study

The aim of this study is to examine recombinant forms of HIV-1 in search for potential new CRFs in Cyprus. The results of this study provide new insights and supply valuable data that aid understanding the influx of new strains that shape the polyphyletic infection of HIV-1 in Cyprus. This study provides the groundwork that facilitate the investigation of new HIV-1 recombinants, consequently aiding in the research of HIV-1, and safeguarding public health.

2.0 MATERIALS AND METHODS

2.1 Methodology Overview

The samples that were used in this study were obtained from HIV-1 serum positive patients diagnosed at the Grigorios HIV Clinic of Larnaca General Hospital. The blood samples were brought to the Laboratory of Biotechnology and Molecular Virology of the University of Cyprus for processing. The serum plasma and PMBCs were initially isolated from whole blood, which were later utilized for HIV-1 RNA and DNA extraction, respectively. The HIV-1 RNA was employed for primary RT-PCR and secondary nested-PCR, and the products were Sanger sequenced for obtaining the pol region (2253-5250 in the HXB2 genome) HIV-1 sequences. The initial analysis was performed on the pol region nucleotide sequences for the identification of drug resistance associated mutations and thus level of resistance to the antiretroviral drugs. This procedure was conducted weekly upon the receival of HIV-1 serum positive samples and the drug resistance reports were sent back to the Grigorios HIV Clinic of Larnaca General Hospital. The genotypic subtypes of the HIV-1 isolates were determined based on the HIV-1 pol region. Subsequently, for near real-time surveillance of the HIV-1 infection and investigation of the HIV-1 transmission dynamics in Cyprus, maximum likelihood phylogenetic analyses were conducted every month. The monthly reports on the growth and characteristics of HIV-1 transmission clusters (molecular clusters) that were identified through the phylogenetic analyses were distributed to public health officers and also posted on the laboratory website (https://www.kostrikislab.com/ongoing-projects/).

Through the monthly phylogenetic analyses that were performed, four transmission clusters of HIV-1 recombinants were identified from the April 2019 report until September 2021 report. The genotypic subtypes of the HIV-1 recombinant were determined based on the pol region sequences. These HIV-1 recombinants were not characterized as previously established CRFs, and hence could be identified as possible new CRFs. The first HIV-1 recombinant transmission cluster, cluster number 4 “Rec. of B, A1, G”, was identified on the April 2019 report, followed by cluster number 5 “Rec. of B, A1” on June 2019 report, and cluster number 9 “Rec. of 02_AG, G, B” along with cluster number 16 “Rec. of B, A1” on August 2019 report. From April 2019 until September 2021 the cluster 4 “Rec. of B, A1, G”, and clusters 5 and 16 “Rec. of B, A1” showed a minor

growth and the cluster 9 “Rec. of 02_AG, G, B” showed an active growth. Specifically, cluster 4 “Rec. of B, A1, G”, cluster 5 “Rec. of B, A1” and cluster 16 “Rec. of B, A1” were increased by one sample, and cluster 9 “Rec. of 02_AG, G, B” was increased by 12 samples.

The aforementioned results led to the hypothesis that those four HIV-1 recombinant transmission clusters were putative CRFs that circulate in Cyprus. Thus, with the aim to perform comprehensive recombination analyses, the near full-length HIV-1 genome sequences of the recombinant samples were obtained, utilizing the three overlapping assays (gag, pol and env amplification and sequencing assays) spanning the whole HIV-1 genome.

The design and development of the assays, including the assay primer design, primary RT-PCR and secondary nested-PCR thermoprofile reactions, as well as the sequencing primers, were created by the members of the Laboratory of Biotechnology and Molecular Virology of University of Cyprus. The pol assay protocol has already been published (Chrysostomou, Topcu et al. 2020); however, it is essential to be described for the methodology of this thesis to be comprehensively explained.

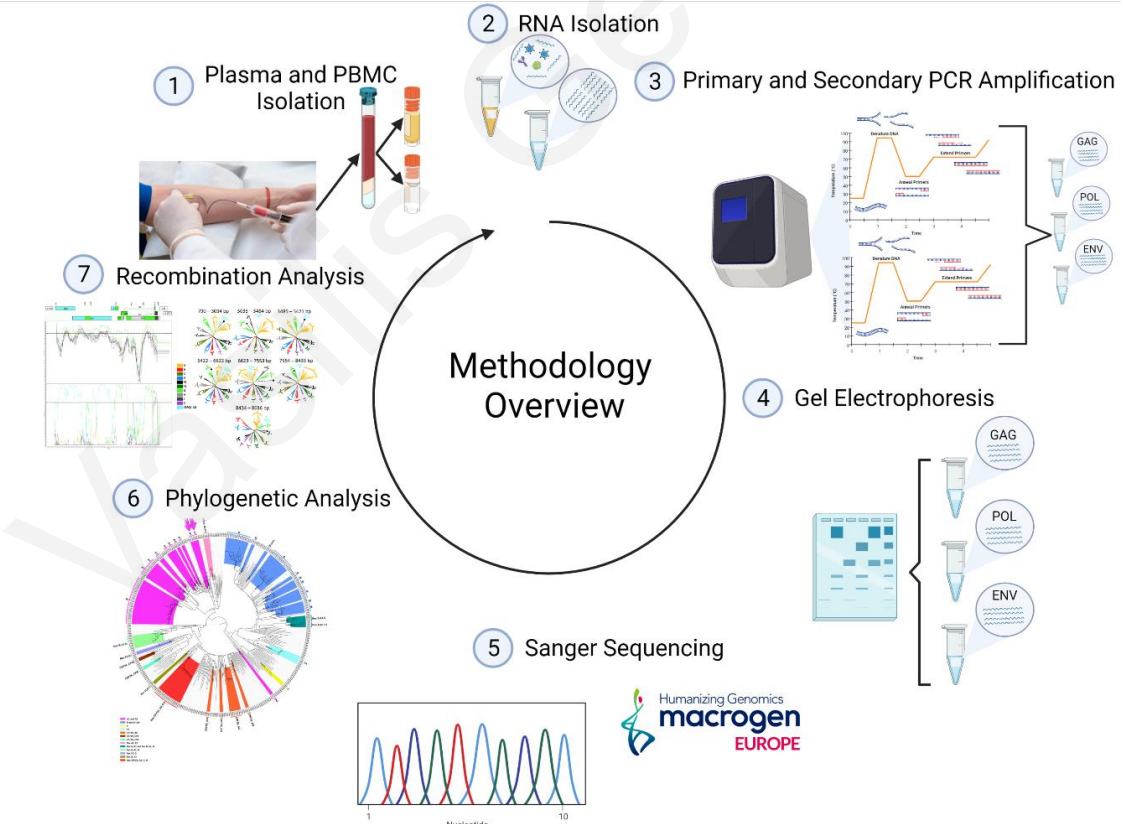


Figure 11: Graphical representation outlining the required methodology conducted during the investigation for potentially new HIV-1 CRFs in Cyprus. The numbers represent the order of the processes used, beginning from the upper left of the schematic representation (1) Plasma and PBMC isolation and ending with (7) the recombination analysis.

2.2 Study Subjects

The blood samples used in the study were obtained from chronic and newly diagnosed HIV-1 serum positive patients diagnosed at the Grigorios HIV Clinic of Larnaca General Hospital. Individuals with a serologically proved diagnosis of HIV infection which prompted the current hospital attendance, or documented during the current hospital stay, were considered newly diagnosed HIV patients (Bishnu, Bandyopadhyay et al. 2014). Patients on HAART were consider as chronic patients (Bishnu, Bandyopadhyay et al. 2014). The samples that were used on the monthly phylogenetic analysis were collected from March 2017 until October 2021. The sampling was implemented following all the regulations of the Cyprus National Bioethics Committee including relevant written informed consent of the study participants for research purposes. An example of the epidemiological data that were provided by patients are demographic information such as gender, age, country of origin and country of infection but also behavioral information such as the route of infection (Table 8 – Table 11).

2.3 Plasma and PBMCs Isolation

Blood samples of patients with a total volume of 8ml were collected by Grigorios HIV Clinic of Larnaca General Hospital in Becton Dickinson Vacutainer® Cell Preparation Tubes (CPT) and sent to the Laboratory of Biotechnology and Molecular Virology at the University of Cyprus. Upon receipt of the samples, were processed with a series of centrifugation for plasma and Peripheral Blood Mononuclear Cells (PBMC) isolation at Biosafety Level 3 facility of the laboratory (Figure 12). Initially, centrifugation was performed at 3000rpm for 20 minutes (eppendorf Centrifuge 5810 R). The plasma layer was then isolated, aliquoted in 500µl volume, placed in Corning® 2mL External Threaded Polypropylene cryogenic vials, and then stored at −80 °C. The layer of PBMCs were then transferred to a 15mL Falcon tube. For the preparation of PBMCs the addition of the appropriate amount of Gibco ™ PBS pH 7.4 was first conducted to make up a total volume of 15mL followed by centrifugation at 1200rpm for 20 minutes. At the end of the centrifugation, the supernatant fraction was discarded while the precipitated fraction was resuspended in 1mL of PBS and transferred to 1.5mL eppendorf tube. This was followed by centrifugation at 1200rpm for 10 minutes. At the end of the centrifugation, the supernatant fraction was removed while the precipitated fraction, the PBMC pellet, was stored at -80 °C.

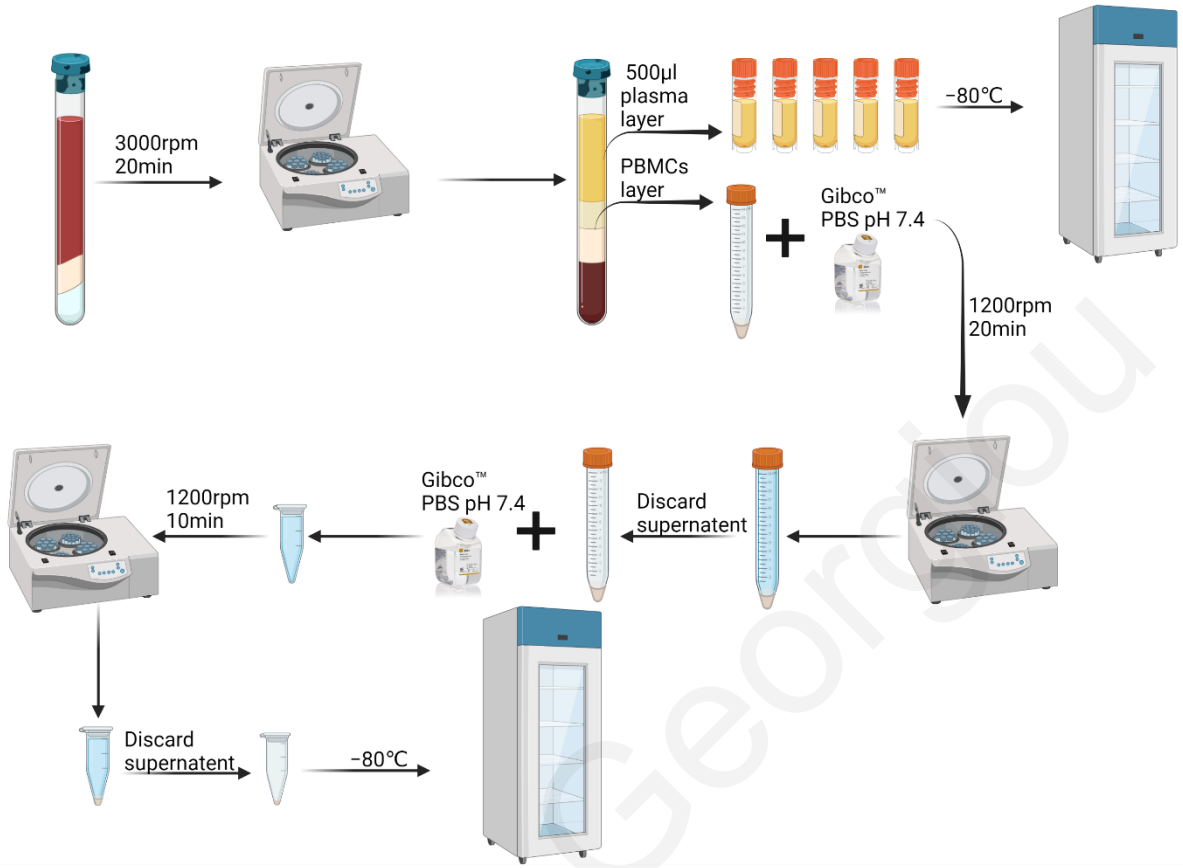


Figure 12: Graphical representation outlining the Plasma and PBMCs isolation.

2.4 RNA Isolation

RNA extraction was performed using a QIAmp Viral RNA Mini Kit (Qiagen, Hilden, Germany) and a QIAcube Connect machine (Qiagen, Hilden, Germany) in accordance with the manufacturer's specifications. For the implementation of the protocol, RNA binding buffers were used as well as carrier, lyse buffer, inhibitor removal buffer, wash buffer and elusion buffers in a series of centrifugations (Figure 13). The result was an elute of 50µl of pure HIV-1 RNA. The isolated RNA was then used for RT-PCR.

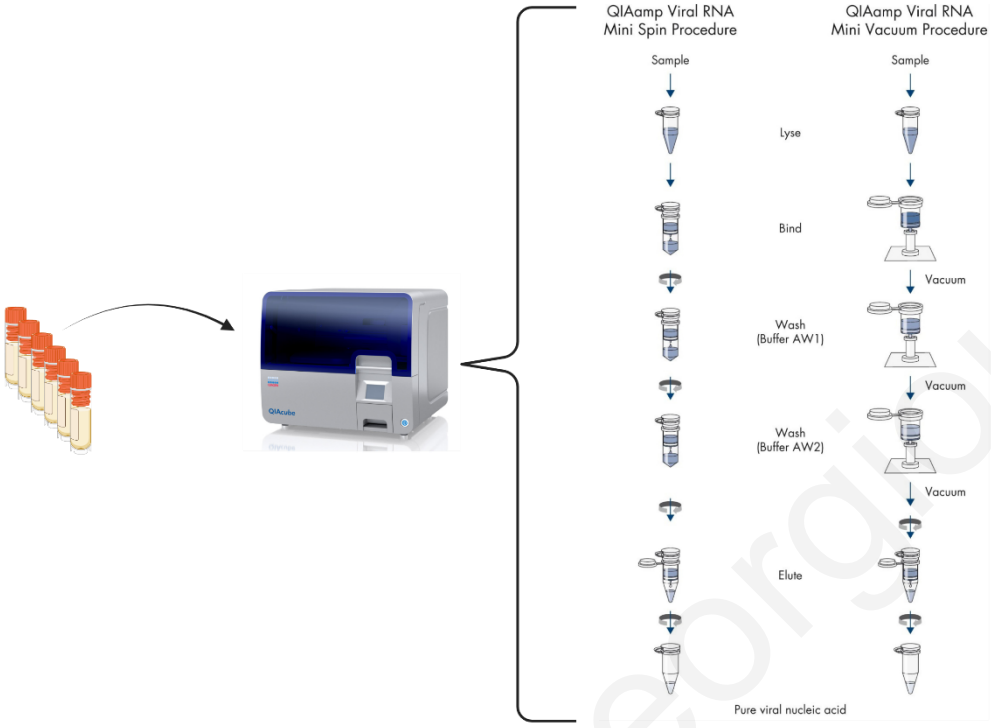


Figure 13: *HIV-1 RNA extraction protocol.* Schematic representation of the steps implemented using the QIAmp Viral RNA Mini Kit (Qiagen, Hilden, Germany) via the QIAcube Connect machine (Qiagen, Hilden, Germany) for the execution of the HIV-1 RNA extraction in accordance with the manufacturer's specifications.

2.5 Primer Development

Most of the primer sequences that were used for the amplification as well for sequencing were designed de novo for this study (Table 1) and were named according to the HXB2 reference genome numbering (GenBank entry number: K03455). The inner and outer PCR primers were precisely designed to target the beginning (forward primers) and the end (reverse primers) of the gag (790 to 2292 in the HXB2 genome), pol (2253-5250 in the HXB2 genome) and env region (5041-8795 in the HXB2 genome) of the HIV-1 genome (Figure 14 B, Table 1). The sequencing primers were designed so that at least two sequencing primers would be aligned at each nucleotide position of the HIV-1 genome (Figure 15 B, Table 1). For the design of the primers, several tools and software have been used. Initially, the use of Geneious® bioinformatics software and ClustalW algorithm for sequence data analysis led to the identification of the appropriate primer binding region. Then for the investigation of the diversity that characterized the primer binding regions, the Quick align tool version 2.0 from the Los Alamos HIV databased was used

(https://www.hiv.lanl.gov/content/sequence/QUICK_ALIGNv2/QuickAlign.html) (Compendium 2018). In cases that a specific position in the primers was characterized by different nucleotides at various subtypes and recombinants of the HIV-1, a wobble was introduced. The term wobble refers to a position in the primer sequence or in the genome that consist isomeric concentrations of different nucleotides. Following that, an analysis of the primers was then performed through Oligo Analyzer tool of the Intergraded DNA Technologies (IDT) (https://eu.idtdna.com/calc/analyzer). The purpose of this analysis was to find the melting temperature (Tm) of the primers as well as the ability of the primers for homo/hetero dimerization. Subsequently, for the introduction of the appropriate temperature of the PCR protocols (Amplification temperature) the annealing temperature of the amplification primers was used. The annealing temperature is the average of the Tm of the two amplification primers subtracting five Celsius degrees. Finally, after the amplification and the sequencing reactions, a primer quality report was created that indicates the performance of the primers (data not shown).

Table 1: Primers that had been used to the primary RT-PCR, secondary (NESTED) PCR and sequencing of gag, pol and env regions.

PCR Primer [a]		Target Gene	Sequence [b]	Position [c]
630	(F)	*GAG*	5'-TAG CAG TGG CGC CC-3'	630 → 643
684	(F)	*GAG*	5'-TCT CGA CGC AGG ACT CG-3'	684 → 705
1810	(F)	*POL*	5'-GCT ACA YTA GAA GAA ATG ATG ACA GCA TG-3'	1810 → 1838
2006	(F)	*POL*	5'-GGG CCC CTA GGA AAA AGG G-3'	2006 → 2024
2404	(R)	*GAG*	5'-CCA ATT CCY CCT ATC ATT TTT GGT TTC C-3'	2377 ← 2404
2501	(R)	*GAG*	5'-GTT GAC AGG TGT AGG TCC TAC-3'	2481 ← 2501
3777	(F)	*ENV*	5'-TGG ATT CCT GAR TGG GAR TTT G-3'	3777 → 3798
4155	(F)	*ENV*	5'-GTA CCA GCA CAC AAA GGR ATT G-3'	4155 → 4180
5264	(R)	*POL*	5'-CCT GTA TGC AGA CCC CAA TAT GTT-3'	5241 ← 5264
5575	(R)	*POL*	5'-TCT GGG GCT TGT TCC ATC TAT C-3'	5554 ← 5575
9038	(R)	*ENV*	5'-TAA GTC ATT GGT CTT ARA GGY ACY TG-3'	9013 ← 9038
9181	(R)	*ENV*	5'-GTG TGT ART TYT GCC AAT CAG G-3'	9160 ← 9181

Seq. Primers [a]		Target Gene	Sequence [b]	Position [c]
684	(F)	*GAG*	5’-TCT CGA CGC AGG ACT CG-3’	684→705
1173	(F)	*GAG*	5’-CAG YCA AAA TTA YCC TAT AGT GCA-3’	1173→1196
1985	(R)	*GAG*	5’-CCT TCY TTG CCA CAR TTG AAA CAY-3’	1962←1985
2404	(R)	*GAG*	5’-CCA ATT CCY CCT ATC ATT TTT GGT TTC C-3’	2377←2406
2454	(F)	*POL*	5’-GGA MAW AAR GCT ATA GGT ACA G-3’	2454→2475
2610	(R)	*POL*	5’-CYT TTG GGC CAT CCA TTC-3’	2593←2610
3003	(F)	*POL*	5’-GGA TGG AAA GGA TCA CC-3’	3003→3019
3019	(R)	*POL*	5’-GGT GAT CCT TTC CAT CC-3’	3003 ← 3019
3462	(R)	*POL*	5’-CTG CCA RTT CTA RYT CTG CTT C-3’	3441 ← 3462
3777	(F)	*POL*	5’-TGG ATT CCT GAR TGG GAR TTT G-3’	3777 → 3798
4060	(R)	*POL*	5’-CCT AAT GCA TAY TGT GAG TCT GTT AC-3’	4035 ← 4060
4155	(F)	*POL*	5’-GTA CCA GCA CAC AAA GGR ATT G-3’	4155 → 4176
4324	(F)	*POL*	5’-TAG CAA AAG AAA TAG TAG CCA GCT G-3’	4324 → 4348
4558	(R)	*POL*	5’-ACT GGC CAT CTT CCT GCT AAT TTT A-3’	4534 ← 4558
4776	(F)	*ENV*	5’-CAC AAT TTT AAA AGA AAA GGG GGG ATT G-3’	4776 → 4803
5554	(F)	*ENV*	5’-GAT AGA TGG AAC AAG CCC CAG A-3’	5554 → 5575
5575	(R)	*ENV*	5’-TCT GGG GCT TGT TCC ATC TAT C-3’	5554 ← 5575
5960	(F)	*ENV*	5’-GGC ATH TCC TAT GGC AGG AAG-3’	5960 → 5980
6203	(F)	*ENV*	5’-GAA AGA GCA GAA GAY AGT GGM A-3’	6203 → 6224
6438	(F)	*ENV*	5’-CAT GCC TGT GTA CCC ACA GA-3’	6438 → 6457
6457	(R)	*ENV*	5’-TCT GTG GGT ACA CAG GCA TG-3’	6438 ← 6457
6858	(F)	*ENV*	5’-CCA ATT CCY ATA CAT TAT TGT GCY C-3’	6858 → 6882
6882	(R)	*ENV*	5’-GRG CAC AAT AAT GTA TRG GAA TTG G-3’	6858 ← 6882
7519	(F)	*ENV*	5’-AAG CAA TGT ATG CCC CTC C-3’	7519 → 7537
8039	(R)	*ENV*	5’-GGT GCA RAT GWG TTT TCC AGA GC-3’	8017 ← 8039
8339	(F)	*ENV*	5’-AAT AGA GTT AGG CAG GGA TAC TCA CC-3’	8340 → 8365
8365	(R)	*ENV*	5’-GGT GAG TAT CCC TGC CTA ACT CTA TT-3’	8340 ← 8365
9011	(R)	*ENV*	5’-GGY CTG ACT GGA AAR CCY AC-3’	8992 ← 9011

[a.] Illustrates the names of the primers that had been used for the primary RT-PCR, secondary (NESTED) PCR, as well as for Sequencing. The direction of the primers is described with the letters F, Forward; R, Reverse.

[b] Demonstrate the 5'- 3' sequences of each primer. In cases that a specific position in the primers is characterized by different nucleotides at various subtypes of the HIV-1, a wobble was introduced. Accordingly, the letter Y refers to equimolar quantities of C and T; R, A and G; M, A and C; W, A and T.

[c] Shows the primer binding site based on the HXB2 numbering (GenBank accession number K03455).

2.6 Primary touchdown Reverse Transcription PCR (RT-PCR)

Upon completion of viral RNA isolation, touchdown primary RT-PCR was performed. At this stage viral RNA was reverse transcribed to complimentary DNA (primary product of RT-PCR). The purpose of the Touchdown approach was to increase the binding specificity of the primers, by the gradual reduction (per reaction cycle) of the initially higher annealing temperature, until it reaches the optimum melting temperature of the primers (Green, Sambrook 2018). The primary touchdown RT-PCR utilized the RNA extracted from HIV-1 plasma to amplify HIV-1 gag (p17, p24, p2, p7, p1, and p6); pol (PR, RT *and* IN); and env (VIF, VPR, TAT, REV, VPU,gp120, gp41) regions, which produced 1871 base pair (bp) product for gag, 3765 bp product for pol and 5404 bp product for env as shown in Figure 14. The primers used for RT-PCR were 630 and 2501 for gag,1810 and 5575 for pol, and 3777 and 9181 for env (Table 1). For the primary RT-PCR, 10 μL of purified HIV-1 RNA from each sample were firstly incubated for 20 seconds at 70 °C to unfold the secondary structure of the RNA, followed by stabilization of RNA by incubation on ice for 1 minute. The rest of the experimental procedure continued to be performed on ice. Subsequently, after the incubation, 40 μL of PCR master mixes were added to each well of the PCR plate containing each RNA sample. The PCR master mixes were previously prepared for each of the three RT-PCR assays. For the preparation of the PCR master mixes, 1 μL of each PCR primer at 20 pmol/μL concentration for gag and pol assays, and 1 μL of each PCR primer at 80 pmol/μL concentration for env assay were added to each of the corresponding master mixes. 25 μL of 2× Platinum SuperFi RT-PCR Master Mix of SuperScript IV One-Step RT-PCR System (ThermoFisher Scientific), 12.5 μL of nuclease free water, and 0.5 μL of the SuperScript IV RT mix were added to make a final volume of 50 μL together with the 10 μL of purified HIV-1 RNA. The amplification primers of the env gene were four times more concentrated (80 pmoles/μl) than the gag and pol amplification primers (20 pmoles/μl) for better reaction efficiency. SuperScript IV One-Step RT-PCR System contains the ezDNase enzyme that acts to remove genomic DNA, which could have been transferred from the plasma HIV-1 RNA isolation step (annex 2.3). PCR

amplifications were carried out in a SimpliAmp™ Thermal Cycler (Life Technologies). All the reagents, their volume and concentrations used for the primary RT-PCR are indicated in the Table 2 below.

Table 2: Components of the Mastermix that had been used for the primary Reverse Transcription PCR (RT-PCR) of gag, pol and env regions

Components	**Quantity**
2X Platinum SuperFi RT-PCR Master Mix	25.0 µl
Nuclease free water	12.5 µl
Outer Forward primer [20µM] or 20 pmoles/µl(*gag* and *pol*) Outer Forward primer [80µM] or 80 pmoles/µl(*env*)	1.0 µl 1.0 µl
Outer Reverse primer [20µM] or 20 pmoles/µl (*gag* and *pol*) Outer Reverse primer [80µM] or 80 pmoles/µl (*env*)	1.0 µl 1.0 µl
SuperScript IV RT mix	0.5 µl
RNA (HIV RNA)	10.0 µl

For the amplification, the thermocycling conditions during the primary touchdown RT-PCR were; one cycle at 50 °C for gag and pol, and 58°C for env, for the reverse transcription reaction for 10 min, followed by one cycle at 98 °C for two minutes for inactivation of the reverse transcriptase and for the initial denaturation of DNA. The env gene the temperature of the initial reaction of the reverse transcription was raised due to the complicated RNA secondary structures (Watts, Dang et al. 2009). Subsequently, the amplification step was followed. The amplification step was divided into two parts, the initial amplification, and the final amplification. The initial amplification was comprised of three phases repeated for 10 cycles. Phase 1 consisted of a 10 s denaturation step at 98 °C. Then, phase 2 consisted of a 10 s annealing step that started at 9 °C higher than the optimal primer melting temperature (T_m) (64°C for gag, 61 °C for pol and 69°C for env), with a ΔT = -1 °C in each following cycle, until the optimal T_m (55°C for gag, 52°C for pol, and 60 for env) was reached at the 10^{th} cycle. Lastly, the phase 3 of the initial amplification step consisted of a two-minute elongation step at 72 °C. The final amplification step of the RT-PCR had the same three phases as the initial amplification with the difference that it consisted of 30 cycles for pol and env and 15 cycles for gag and the annealing temperature remained at the optimal T_m according to the primer set of each gene. The final elongation step was performed at 72 °C for five minutes. All the aforementioned PCR thermal cycling conditions are summarized in the Table 3 below.

Table 3: Primary touchdown RT-PCR conditions for gag, pol and env regions [a]			
Step [c]	**Temperature (°C)**	**Time (min)**	**Number of Cycles**
Reverse transcription	50 (gag and pol) 58 (env)	10:00	1×
RT Inactivation/initial denaturation	98	2:00	1×
Initial amplification	98	0:10	10x
	64-55 (gag) 61–52 (pol) 69-60 (env) (ΔT = -1 °C) [b]	0:10	
	72	1:00 (gag) 2:00 (pol) 3:00 (env)	
Final amplification	98	0:10	15x (gag) 30x (pol and env)
	55 (gag) 52 (pol) 60 (env)	0:10	
	72	1:00 (*gag*) 2:00 (*pol*) 3:00 (*env*)	
Final extension	72	5:00	1×
Reaction stop	4	Indefinitely	Hold

[a] Thermocycling conditions of the Reverse Transcription PCR that lead to the production of the cDNA for each gene.

[b] The initial amplification was based on a touchdown approach that starts with 9 °C higher than the melting Temperature (Tm) and proceeds with ΔT = −1 °C for 10 cycles. Then it reaches the annealing temperature and stays stable.

[c] The step that leads to each stage of the Polymerase Chain Reaction

2.7 Secondary (NESTED) PCR

After the primary RT-PCR, a secondary (NESTED) PCR was performed. At the secondary (NESTED) PCR the amplicons produced from the primary RT-PCR reaction were used as a template for a second amplification by using a second set of primers. As a result, the sensitivity and specificity of DNA amplification was significantly enhanced (Carr, Williams et al. 2010). The primers for the nested PCR were 684 and 2404 for gag, 2006 and 5264 for pol and 4155 and 9038 for env (Table 1). Subsequently, 45 μL of PCR master mixes were added to each well of the PCR plate containing each DNA sample. The PCR master mixes were previously prepared for each of the three secondary (NESTED) PCR assays. For the preparation of the PCR master mixes, 1 μL of each PCR primer at 20 pmol/μL concentration for gag and pol assays, and 1 μL of the forward PCR primer at 40 pmol/μL concentration along with 1 μL of the reverse PCR primer at 80 pmol/μL concentration for env assay were added to each of the corresponding master mixes. 25 μL of 2× Platinum™ Hot Start PCR Master Mix (ThermoFisher Scientific) for gag and pol, 25 μL of Invitrogen™ Platinum™ SuperFi™II PCR Master Mix (ThermoFisher Scientific) for env, and 18 μL of nuclease free water were added to make a final volume of 50 μL together with the 5 μL of purified HIV-1 DNA. The different DNA polymerase enzyme that was used for the env gene it was due to the larger size of the final amplification product. The amplification primers of the env gene were more concentrated (forward: 40 pmoles/μl, reverse: 80 pmoles/μl) than the gag and pol amplification primers (20 pmoles/μl) for better reaction efficiency. PCR amplifications were carried out in a SimpliAmp™ Thermal Cycler (Life Technologies). All the reagents, their volume and concentrations used for the secondary (NESTED) PCR are indicated in Table 4 and Table 5 below.

Table 4: Components of the Mastermix that had been used for the secondary (NESTED) PCR of gag and pol regions

Components	Quantity
2X Platinum™ Hot Start PCR Master Mix	25.0 μl
Nuclease free water	18.0 μl
Inner Forward primer [20μM] or 20 pmoles/μl	1.0 μl
Inner Reverse primer [20μM] or 20 pmoles/μl	1.0 μl
RT-PCR Product	5.0 μl

Table 5: Components of the Mastermix that had been used for the secondary (NESTED) PCR of env region

Components	Quantity
Invitrogen™ Platinum™ SuperFi™II PCR Master Mix	25.0 µl
Nuclease free water	18.0 µl
Inner Forward primer [40µM] or 40 pmoles/µl	1.0 µl
Inner Reverse primer [80µM] or 80 pmoles/µl	1.0 µl
RT-PCR Product	5.0 µl

The first step of the nested touchdown PCR was the denaturation of DNA and activation of Platinum *Taq* DNA polymerase for gag and pol that was implemented with a two-minute cycle at 94 °C. The amplification step was similar as the primary RT-PCR. The initial amplification starts with the phase 1 that was the denaturation, performed for 30 s at 94 °C. Then the Phase 2 consisted of a 30 s annealing step that started at 64°C for gag and 62 °C for pol, with a $\Delta T = -1$ °C in each following cycle, for 10 cycles until it reaches the optimal Tm that is 55°C for gag and 53 °C for pol. Phase 3 of the initial amplification step consisted of one minute for the gag and a three and a half minutes for pol, elongation step at 72 °C. The final amplification step of the secondary (NESTED) PCR consisted of 30 cycles of the matching three phases as the initial amplification step, but at the amplification stage, the temperature remained at the optimal Tm of the gag (55°C) and pol (53 °C). The final step was an elongation step at 72 °C for five minutes

For the env gene different thermocycling conditions were applied due to the usage of different enzyme. The initial step was the denaturation of DNA and activation of Platinum SuperFi II DNA Polymerase that was implemented at 98°C for 30 seconds. Then a single amplification step for 40 cycles was implemented with the denaturation occurred on phase 1, performed for 30 s at 98 °C, followed by Phase 2 consisted of a 10 s annealing step at 60°C leading to the phase 3 consisted of three-minute elongation step at 72 °C. The final step was an elongation step at 72 °C for five minutes. All the aforementioned PCR thermal cycling conditions are summarized in and the Table 6 and Table 7 below.

Table 6: Secondary (NESTED) touchdown PCR conditions for gag and pol regions [a]			
Step [c]	**Temperature (°C)**	**Time (min)**	**Number of Cycles**
Initial denaturation/tag *pol*ymerase activation	94	2:00	1×
Initial amplification	94	0:30	10×
	64-55 (gag) 62–53 (pol) (ΔT = −1 °C) [b]	0:30	
	72	1:00 (gag) 3:30 (pol)	
Final amplification	94	0:30	30×
	55 (gag) 53 (pol)	0:30	
	72	1:00 (gag) 3:30 (pol)	
Final extension	72	5:00	1×
Reaction stop	4	Indefinitely	Hold

Table 7: Secondary (NESTED) PCR conditions for env region [a]			
Step [c]	**Temperature** [°C]	**Time [min]**	**No. Cycles**
Initial denaturation	98	0:30	1X
Amplification	98	0:10	40X
	60	0:10	
	72	3:00	
Final Extension	72	5:00	1X
Reaction Stop	4	Indefinitely	Hold

[a] Thermocycling conditions of the secondary NESTED PCR for the amplification of the Primary RT-PCR product.

[b] The initial amplification was based on a touchdown approach that starts with 9 °C higher than the melting Temperature (Tm) and proceeds with ΔT = −1 °C for 10 cycles. Then it reaches the Tm and stays stable.

[c] The step that leads to each stage of the Polymerase Chain Reaction

The conditions for both Primary RT-PCR and secondary (NESTED) PCR were modified according to the size of the PCR product (gag, pol and env regions), the Tm of the primers, the RNA secondary structure of each region and the manufacturer's instructions. The primers used as well as the final products produced at the end of the amplification and the sequencing processes are shown in Figure 14 and Figure 15.

2.8 Gel Electrophoresis

Next, electrophoresis was performed to evaluate the results of the PCR assays on 1% agarose gel. To make the gel, 1.5g of agarose was liquefied with the addition of 150ml Tris-Acetate-EDTA (TAE) buffer. Then the homogenization of the liquid was performed through heating followed by the addition of 7.5μl of GelRed® Nucleic Acid Gel Stain (GelRed). GelRed has the ability to penetrate between the nucleotide bases of the genetic material, thus allowing their detection under ultraviolet radiation (UV). The liquefied 1% agarose was then solidified in a mold provided by the electrophoresis device. After the stabilization of the gel 9μl of the final product of secondary (NESTED) PCR (for each sample) and the controls, together with 1μl 10X Loading Buffer (Takara Bio Inc), were loaded on the gel followed by the loading of the 1kb DNA ladder (Nippon Genetics Dueren, Germany). The controls used are the positive controls and the blank. The positive controls refer to known positive RNA and DNA samples, which indicate the validation of the applied methodologies. The blank control, contained all the reagents except the HIV-1 samples, and it was used for the validation in terms of contamination. The electrophoresis conditions were programmed for one hour on 100V, 100 mA and 100 W. At the end of electrophoresis, the gel was transferred to the BioImaging System GeneGenius (SynGene, Cambridge, United Kingdom), which by producing UV radiation combined with the use of GeneSnap software, enabled the visual display of the secondary (NESTED) PCR product on the gel (Figure 17, Figure 20 and Figure 21). The amplified products from the secondary PCR were sent to Macrogen Europe (https://dna.macrogen-europe.com/eng/member/login.jsp) for purification and Sanger sequencing.

2.9 Sequencing and Genome Analysis

The positive results together with the sequencing primers were then sent to Macrogen (Macrogen-Europe) biotechnology company for Sanger Sequencing. The sequencing of each region (gag, pol and env) was implemented independently. The sequencing results were provided by Macrogen in the form of primer chromatographs for each sample separately. For the visualization, analysis, and quality control of the results, the Geneious software was used (Kearse, Moir et al. 2012). Through the Geneious software the primer chromatographs for each sample were mapped and aligned with the HXB2 reference genome (GenBank accession number K03455), resulting in the consensus sequence of the gag, pol and env regions. The validation of the results were based on the quality of the primer chromatographs, the depth of the primer coverage and the comparison of the sample consensus with the HXB2 reference genome in each nucleotide position separately (Figure 16). Additionally, when the signal of the primers were characterized by overlapping peaks (or when one signal was higher than 50% of the other) then that position in the genome consisted of isomeric concentrations of different nucleotides (termed as wobble) and therefore from a high representation of different quasi-species (Domingo, Escarmís et al. 2008) (Figure 16).

In the second phase of the genome analysis, the completed pol (2253-5250 on HXB2 genome) region sequences were analyzed through Stanford University HIV-1 Drug Resistance Database (Tang, Liu et al. 2012) to find antiretroviral resistance associated mutations (Figure 18). Stanford University HIV-1 Drug Resistance Database is a database located at Stanford University that monitors 200 mutation patterns categorized as primary and polymorphic accessory mutations of HIV-1 that enable the virus to have high or low resistance to specific antiretroviral drugs accordingly (Bennett, Camacho et al. 2009).

The third phase of the genome analysis was the identification of the HIV-1 genotypic subtypes that characterize each sample. For this procedure all the pol region (2253-5250 on HXB2 genome) sequences were analyzed using the REGA HIV-1 Subtyping Tool version 3 (Pineda-Peña, Faria et al. 2013) (Table 12 - Table 15). The REGA HIV-1 Subtyping Tool use phylogenetic methods to detect the subtype of a particular sequence and bootscan methods for the recombination analysis.

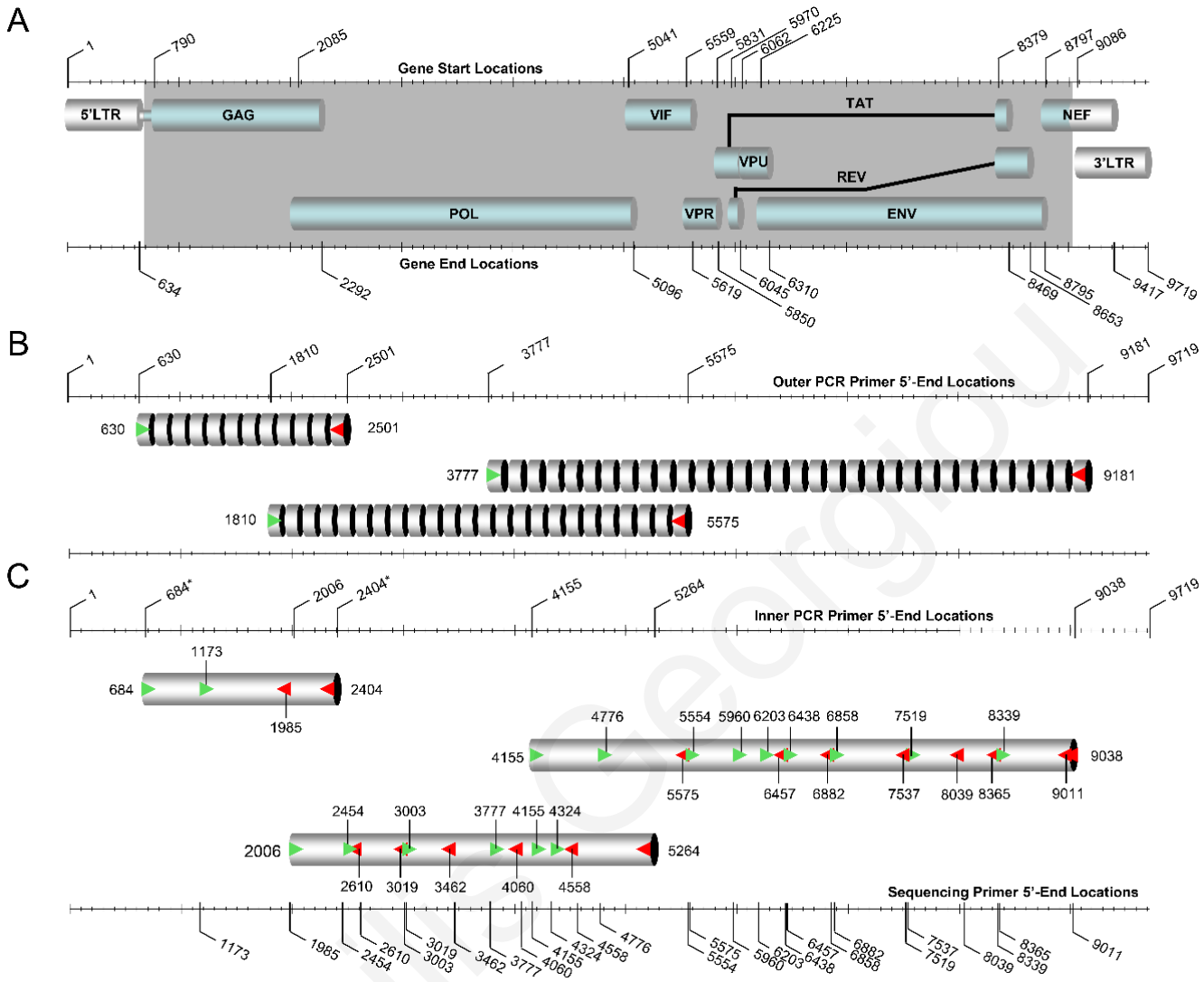


Figure 14: *Graphical Representation of the PCR and Sequencing Designed.* The schematic illustration reviews the amplification and sequencing design of the near full genome of HIV-1 M Group subtypes, circulating recombinant forms (CRFs) and Unique recombinant forms (URF). (A) The top part of the figure illustrates the overall genetic organization of the HIV-1 based on the HXB2 reference genome (GenBank accession number K03455), where each major genetic region is presented as labelled cylindrical figure. The respective numbers above and below the diagram indicate the corresponding beginning and end of each gene with respect to the HXB2 genome. The gray shaded area, on the HIV-1 gene map, represents the region of amplification and sequencing. (B) The middle part of the figure represents the primary RT-PCR design where each set of the amplification primers are represented at the beginning and the end of each correspondence gene beginning with the gag (630 and 2501), pol (1810 and 5575) and env (3777 and 9181). The direction of the primers is indicated with the color (Green: Forward; Red: Reverse) and the arrows. The respective numbers above indicate the 5′-end-primer-binding positions based on the HXB2 numbering. (C) The bottom part of the figure demonstrates the secondary (NESTED) PCR and sequencing primer design. The set of the amplification primers are represented at the beginning and the end of each correspondence gene beginning with the gag (684 and 2404), pol (2006 and 5264) and env (4155 and 9038). The rest of the primers illustrated in between of the amplification primers represent the sequencing primers of each gene. The respective numbers on the scale above and below indicate the 5′-end-primer-binding positions based on the HXB2 numbering. The direction of the primers is indicated with the color (Green: Forward; Red: Reverse) and the arrows.

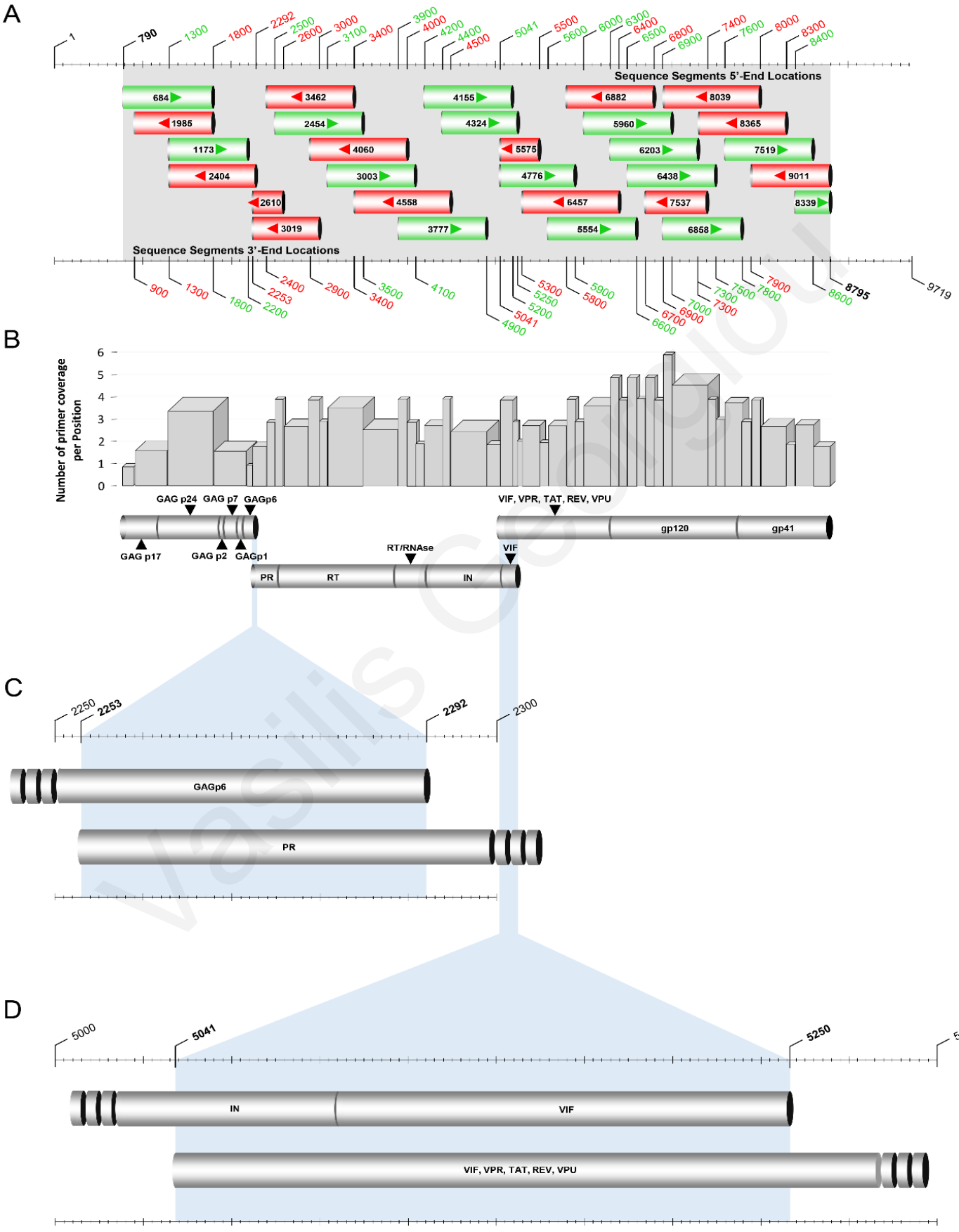

A
Sequence Segments 5'-End Locations
Sequence Segments 3'-End Locations
B
Number of primer coverage per Position
GAG p24
GAG p7
GAGp6
GAG p17
GAG p2
GAGp1
RT/RNAse
VIF
VIF, VPR, TAT, REV, VPU
gp120
gp41
PR
RT
IN
C
GAGp6
PR
D
IN
VIF
VIF, VPR, TAT, REV, VPU

Figure 15: *Sequencing design of the near full*-length *genome of the HIV-1*. (A) The green and red cylindrical figures represent the sequence segments acquired from the correspondingly labeled sequencing primers. The orientations of these sequencing primers are indicated with the equivalent color (Green: Forward; Red: Reverse) and arrows. The respective numbers above and below the diagram indicate the beginning and end of the sequence segments in accordance with HXB2 numbering. (B) The solid cylindrical figures indicate the resulting DNA sequence created of all the partial segments (corresponding to nucleotides 790 to 8795 on the HXB2 genome). This final aligned DNA sequence contains the full gag (nucleotide position 790 to 2292), full pol (nucleotide position 2085 to 5096), VIF (nucleotide position 5041 to 5619), VPR (nucleotide position 5559 to 5850), VPU (nucleotide position 6062 to 6310), TAT (nucleotide position 5831 to 6045 and 8379 to 8469), REV (nucleotide position 5970 to 6045 and 8379 to 8653) and full env (nucleotide position 6225 to 8795) genes. The graph above of the resulting aligned DNA sequence demonstrates the number of aligned nucleotides per position (primer coverage) when all the partial segments are composed together. (C) Indicate the sequenced overlap region between the gag and the pol region. The scale above the cylinders shows the overlap region's beginning (2253) and the end (2292). (D) Indicate the sequenced overlap region between the pol and the env region. The scale above the cylinders shows the overlap region's beginning (5041) and the end (5250).

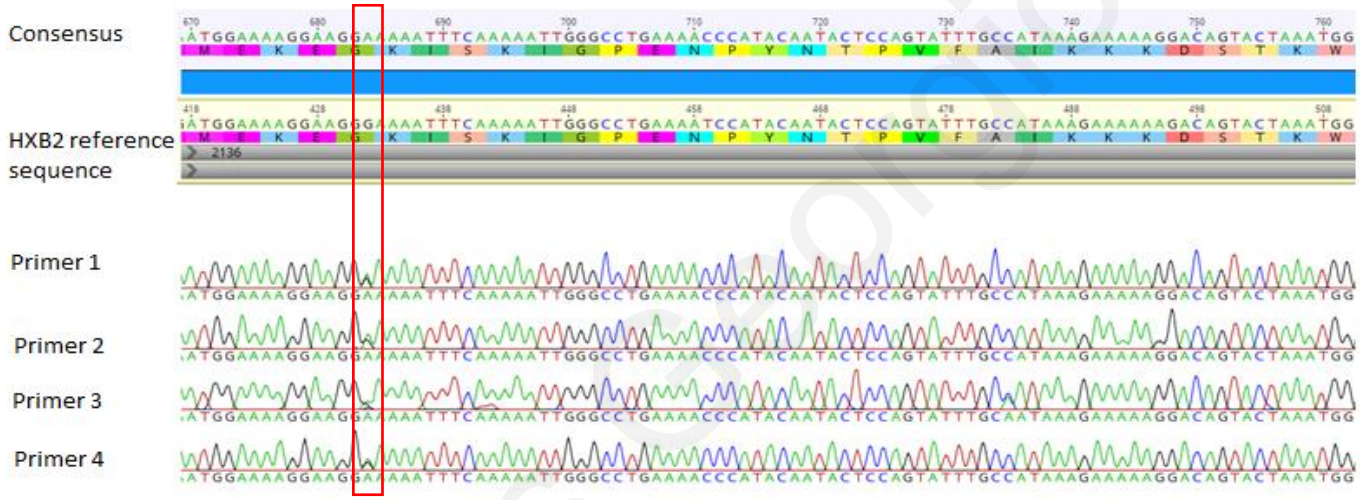


Figure 16: Example of the sequencing analysis through the Geneious software (Kearse, Moir et al. 2012)**.** The figure represents the alignment of the primers for the creation of the consensus sequence of the pol region. The top section of the figure illustrates the consensus nucleoside and amino acid sequence of the HIV-1 positive sample, followed by the HXB2 nucleoside and amino acid sequence. The middle to bottom section of the figure represents the alignment of the chromatographs of four independent primers (primer 1- primer 4). The quality control of the sequence was achieved by analyzing the peaks of the primer chromatographs, the depth of the primer coverage and the comparison of the sample consensus with the HXB2 reference genome in each nucleotide position separately. The red box represents an example of a wobble in which the signal of all four primers was characterized by overlapping peaks resulting in the nucleotide at position 432 being presented as either A or G. The figure is a part of the sequencing result of the pol gene from unpublish data.

2.10 Phylogenetic Analysis

Following the drug resistance analyses and the subtype determination, the data were introduced into the MEGA-X software for the purpose of phylogenetic analysis (Kumar, Stecher et al. 2018). Through the MEGA-X software, the ClustalW algorithm was implemented to align the sequences of interest with reference sequences. The ClustalW algorithm is a multiple sequence alignment (MSA) methodology and works by using groups of similar sequences, with the ability to repeat the alignment process, until a stable alignment is obtained (Feng, Da-Fei, Doolittle 1987, Thompson, Higgins et al. 1994, Wiley, Lieberman 2011). Then, the alignments were visually inspected and manually edited with the AliView v.1.26 algorithm (Larsson 2014). Editing allowed the removal of remaining putative sequencing artifacts (i.e., variations adjacent to gap regions that were not shared with other sequences). The algorithms used to create phylogenetic trees were the maximum-likelihood and the neighbor-joining. Creating maximum-likelihood trees through the MEGA-X software was performed with nucleotide substitution, General Time Reversible model (GTR) and with bootstrap value of 1,000 replicates. The bootstrap of 1,000 repeats was used for the assessing the reliability of the results. The nucleotide substitution model GTR for all sequencing data was calculated by the MEGA-X software. The creation of phylogenetic trees through the neighbor-joining algorithm was performed with nucleotide substitution, Kimura 2-parameter model and with bootstrap value of 1,000 replicates. The purpose of the phylogenetic analysis was the identification and visualization of the genetic similarity among the different HIV-1 strains (Lam, Hon et al. 2010). The correlation between the different strains of HIV-1 has become more obvious by using the Cluster Picker program for phylogenetic clustering analyses (Ragonnet-Cronin, Hodcroft et al. 2013). The parameters used in Cluster Picker were initial threshold 0.7, main support threshold 0.7, genetic distance threshold 4.5, and large cluster threshold 40. For visualization and esthetic manipulation of the trees the FigTree version 1.4.4 was used (http://tree.bio.ed.ac.uk/software/figtree/). All values and parameters used were obtained from the relative literature on the phylogenetic analysis of HIV-1 strains (Pineda-Peña, Schrooten et al. 2014, Pineda-Peña, Theys et al. 2018). Due to the extensive need for the construction of maximum likelihood trees and to alleviate the computational burden, near full-length HIV-1 genomes were multiple aligned using ClustalW algorithm in MAFFT v.7.475 multiple sequence alignment program (Katoh, Standley 2013, Katoh, Rozewicki et al. 2019). The maximum likelihood trees

were subsequently created by using IQTREE v.2.1.2 software (Nguyen, Schmidt et al. 2015, Trifinopoulos, Nguyen et al. 2016), with branch support estimated through the SH-like approximate likelihood ratio test (SH-aLRT) (Guindon, Dufayard et al. 2010) and ultrafast bootstrap analysis (UFBoot) (Minh, Nguyen et al. 2013).

2.11 Recombination Analysis

The near full-length HIV-1 sequences of 24 possible HIV-1 recombinant samples were obtained. To investigate the mosaic pattern of the near full genomes in addition to the analysis performed by the REGA HIV-1 Subtyping Tool (Pineda-Peña, Faria et al. 2013) for the pol gene, a bootscan and a similarity plot analysis were performed using the Simplot, version 3.5.1 software (Lole, Bollinger et al. 1999). This approach allows the identification and evaluation of potential genome breakpoints as well as the combination of subtypes that characterize the sequence through graphical detection of changes in phylogenetic signal. The bootscan and similarity analyses was performed using sliding window of 400 nucleotides overlapping by 40 nucleotides. The parameters used were based on a related study conducted by Kousiappa, Van De Vijver and Kostrikis at 2009 (Kousiappa, Van De Vijver, David AMC et al. 2009).

This was followed by the implementation of the jumping profile hidden Markov model (jpHMM) (Schultz, Bulla et al. 2012), which predicts whether the query sequence is a result of recombination event and in addition allows the accurate identification of the breakpoints and the subtypes characterizing the query sequence. The results of the jpHMM analysis indicate the recombination pattern, presenting it both graphically and with a table that lists the coordinates and subtypes that characterize each section of the genome separately.

Next, for further analysis and evaluation of the results provided by the Simplot software and the jpHMM method, the samples were analyzed using the Recombinant Identification Program (RIP), which is a tool of the Los Alamos international HIV sequence database (www.hiv.lanl.gov/). RIP is a program designed to screen rapidly for HIV type 1 inter-subtype recombinant sequences (SIEPEL, HALPERN et al. 1995).

In addition to the previous approaches another similarity-based method was implemented through the NCBI genotyping tool (https://www.ncbi.nlm.nih.gov/projects/genotyping/formpage.cgi), a web-based genotyping resource for viral sequence. The parameters used on the NCBI genotyping tool were the same with

the Simplot, version 3.5.1 software. This approach was applied for even further validation of the results provided by the previous approaches but also due to the easiness that provides in terms of the manipulation of the reference dataset.

For the verification and validity of the results only boot-scanning and similarity methodologies are not sufficient but further confirmation is required using appropriate phylogenetic methods (Rozanov, Plikat et al. 2004, Gale, Myers et al. 2004). Thus, the construction of neighbor-joining trees with the Kimura two-parameter method and 1,000 repetitions using the MEGA-X software was implied (Kumar, Stecher et al. 2018) for each segment of the genome separately. For the construction of neighbor-joining trees, as well as for the similarly and bootscan analysis approaches, the reference genomes of all pure subtypes of HIV-1 group M were obtain through the Los Alamos Rip alignment of 2017 and 2020. In particular, the pure subtypes of the RIP 2017 alignment were used for the cluster 9 (CRF91_cpx and URF of CRF91_cpx, B) recombination analysis (due to the availability at the time), and the pure subtypes of the RIP 2020 alignment were used for the recombination analysis of the other three putative CRFs. The CRF02_AG, CRF06_cpx, CRF56_cpx and several B subtype sequences that were included into the reference dataset were obtained through Basic Local Alignment Search Tool (blast) analysis (Table S1-S4).

To verify the uniqueness of the recombinant genomes the construction of Maximum Likelihood trees was performed by using IQTREE v.2.1.2 software (Nguyen, Schmidt et al. 2015, Trifinopoulos, Nguyen et al. 2016), with branch support estimated through the SH-like approximate likelihood ratio test (SH-aLRT) (Guindon, Dufayard et al. 2010) and ultrafast bootstrap analysis (Minh, Nguyen et al. 2013). Tree construction was performed using reference samples from all subtypes and CRFs obtained from the RIP Alignment dataset of 2020 of the international Los Alamos HIV sequence database (Table S1-S4). Finally, Blastn was implemented through the NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and the Los Alamos database (https://www.hiv.lanl.gov/content/sequence/BASIC_BLAST/basic_blast.html) in order to find possible identical sequences.

3.0 RESULTS

3.1 Epidemiological Information of the Study Subjects

For the implementation of this research 25 seropositive HIV-1 patients were analyzed. All patients fulfilled and signed their written consent and procedures were applied according to the regulations of the Cyprus National Bioethics Committee. The study subjects were initially analyzed based on the pol gene (annex 3.2) as part of the near real-time surveillance of drug resistance mutation project of the Laboratory of Biotechnology and Molecular Virology (BMV) of University of Cyprus. For the investigation of drug resistance mutations as well as for the investigation and characterization of the HIV-1 transmission dynamics in Cyprus, monthly maximum likelihood trees were created and uploaded on the BMV lab website (https://www.kostrikislab.com/ongoing-projects/). Based on the initial phylogenetic analysis of the pol region, among all the samples collected from March 2017 until October 2021, 31 samples were clustered in 4 different recombinant clusters. From those 31 samples the near full genome of 24 samples was successfully obtained, one was partially amplified and sequenced (CY590 gag and pol region) and two samples were added into the recombinant cluster 9 "Rec. of 02_AG, G, B" after the analysis. The remaining four samples failed. The epidemiological information of the patients that were included on this study are indicated at the tables below divided according to the appropriate clusters.

Table 8: Clinical and epidemiological information for study patients of the cluster 9"Rec. of 02_AG, G, B" (CRF91_cpx)

Patient [a]	Sex [b]	Age (yr)	Positive Test Date [c]	Country of origin [d]	Risk group [e]	CD4 (cells/mm³)	Plasma HIV-1 RNA (copies x 10^4/ml)	Epidemiological information [f]
CY467	M	33	06/17	Nigeria	HBC	345	7,99	Infected in Nigeria
CY494	M	31	01/18	Cyprus	HC	310	21,70	Infected in Cyprus
CY520	M	49	03/18	Cyprus	MSM	300	28,70	Infected in Cyprus
CY533	M	34	05/18	Cyprus	HC	851	1,59	Infected in Cyprus
CY614	M	28	04/19	Cyprus	MSM	237	126,00	Infected in Cyprus
CY622	M	69	05/19	UK	MSM	694	8,78	N/A
CY630	M	74	-/19	Cyprus	MSM	294	32,40	Infected in Cyprus
CY640	M	41	09/19	Cyprus	MSM	426	31,70	Infected in Cyprus
CY670	M	58	01/20	Bulgaria	MSM	103	46,30	Infected in Cyprus
CY686	M	60	05/20	UK	MSM	131	107,00	Infected in Cyprus
CY742	M	45	09/20	Cyprus	HBC	309	86,40	Infected in Cyprus

[a] Indicates the laboratory code for each study subject. [b] M, male. [c] Indicates the date (month/year; -, unknown month) of the first known positive HIV antibody test. [d] Country of birth of the study subjects; UK, United Kingdom. [e] HBC, hetero-bisexual contact; MSM, men who have sex with men; HC, heterosexual contact; [f] Information provided by the study subjects. N/A, Not available.

Table 9: Clinical and epidemiological information for study patients of the Cluster 4 "Rec. of B, A1, G"

Patient [a]	Sex [b]	Age (yr)	Positive Test Date [c]	Country of origin [d]	Risk group [e]	CD4 (cells/mm³)	Plasma HIV-1 RNA (copies x 10^4/ml)	Epidemiological information [f]
CY448	M	38	08/15	Cyprus	MSM	739	3,08	Infected in Cyprus
CY512	M	59	03/18	Cyprus	MSM	150	0,03	Infected in Cyprus
CY525	M	56	04/18	Cyprus	MSM	124	13,60	Infected in Greece
CY526	M	28	04/18	Lebanon	MSM	631	0,70	Infected in Cyprus
CY529	M	29	05/18	Greece	HBC	357	7,66	Infected in Greece
CY537	M	55	06/18	Cyprus	HC	729	14,60	Infected in Cyprus
CY625	M	34	07/19	Cyprus	HBC	406	26,90	Infected in Cyprus

[a] Indicates the laboratory code for each study subject. [b] M, male. [c] Indicates the date (month/year; -, unknown month) of the first known positive HIV antibody test. [d] Country of birth of the study subjects; [e] HBC, hetero-bisexual contact; MSM, men who have sex with men; HC, heterosexual contact; [f] Information provided by the study subjects.

Table 10: Clinical and epidemiological information for study patients of the cluster 5 "Rec. of B, A1"

Patient [a]	Sex [b]	Age (yr)	Positive Test Date [c]	Country of origin [d]	Risk group [e]	CD4 (cells/mm³)	Plasma HIV-1 RNA (copies x 10^4/ml)	Epidemiological information [f]
CY397	M	35	N/A	Cyprus	N/A	N/A	N/A	N/A
CY413	M	54	10/13	Cyprus	HBC	1216	2,50	Infected in Cyprus
CY590	M	36	05/12	Cyprus	MSM	229	2,97	Infected in Cyprus

[a] Indicates the laboratory code for each study subject. [b] M, male. [c] Indicates the date (month/year; -, unknown month) of the first known positive HIV antibody test. [d] Country of birth of the study subjects; [e] HBC, hetero-bisexual contact; MSM, men who have sex with men; [f] Information provided by the study subjects. N/A, Not available.

Table 11: Clinical and epidemiological information for study patients of the cluster 16 "Rec. of B, A1"

Patient [a]	Sex [b]	Age (yr)	Positive Test Date [c]	Country of origin [d]	Risk group [e]	CD4 (cells/mm³)	Plasma HIV-1 RNA (copies x 10^4/ml)	Epidemiological information [f]
CY584	M	54	01/19	Cyprus	MSM	9	0,38	Infected in Venezouela
CY620	M	47	05/19	Cyprus	HC	450	12,40	Infected in Cyprus
CY697	M	34	04/20	Cyprus	MSM	500	7,31	Infected in Cyprus
CY710	M	39	06/20	Romania	HBC	330	5,51	Infected in Romania

[a] Indicates the laboratory code for each study subject. [b] M, male. [c] Indicates the date (month/year; -, unknown month) of the first known positive HIV antibody test. [d] Country of birth of the study subjects; [e] HBC, hetero-bisexual contact; MSM, men who have sex with men; HC, heterosexual contact; [f] Information provided by the study subjects.

3.2 Amplification and Sequencing of the pol region

Upon receipt of the seropositive HIV-1 samples from Grigorios HIV-1 Clinic of Larnaca General Hospital an analysis regarding the pol region (2253-5250 in the HXB2 genome) was performed for the identification of possible antiretroviral resistance mutations and thus resistance to the antiretroviral therapy. To accomplish this purpose the previously described methodologies were performed, including the Plasma and PBMCs isolation (annex 2.3), RNA isolation (annex 2.4) Primary touchdown Reverse Transcription PCR (annex 2.6), Secondary (NESTED) PCR (annex 2.7) and Gel Electrophoresis (annex 2.8). Based on the amplification primers of the secondary (NESTED) PCR (Table 1) the expected amplified fragment of the HIV-1 was 2998 bp and as such on the gel images is represented at approximately 3000 bp according to the 1Kb DNA ladder. The positive samples were then sent for Sanger sequencing to Macrogen (Macrogen-Europe). Subsequently, the quality of the sequencing results were evaluated using Geneious software (Kearse, Moir et al. 2012) (annex 2.9).

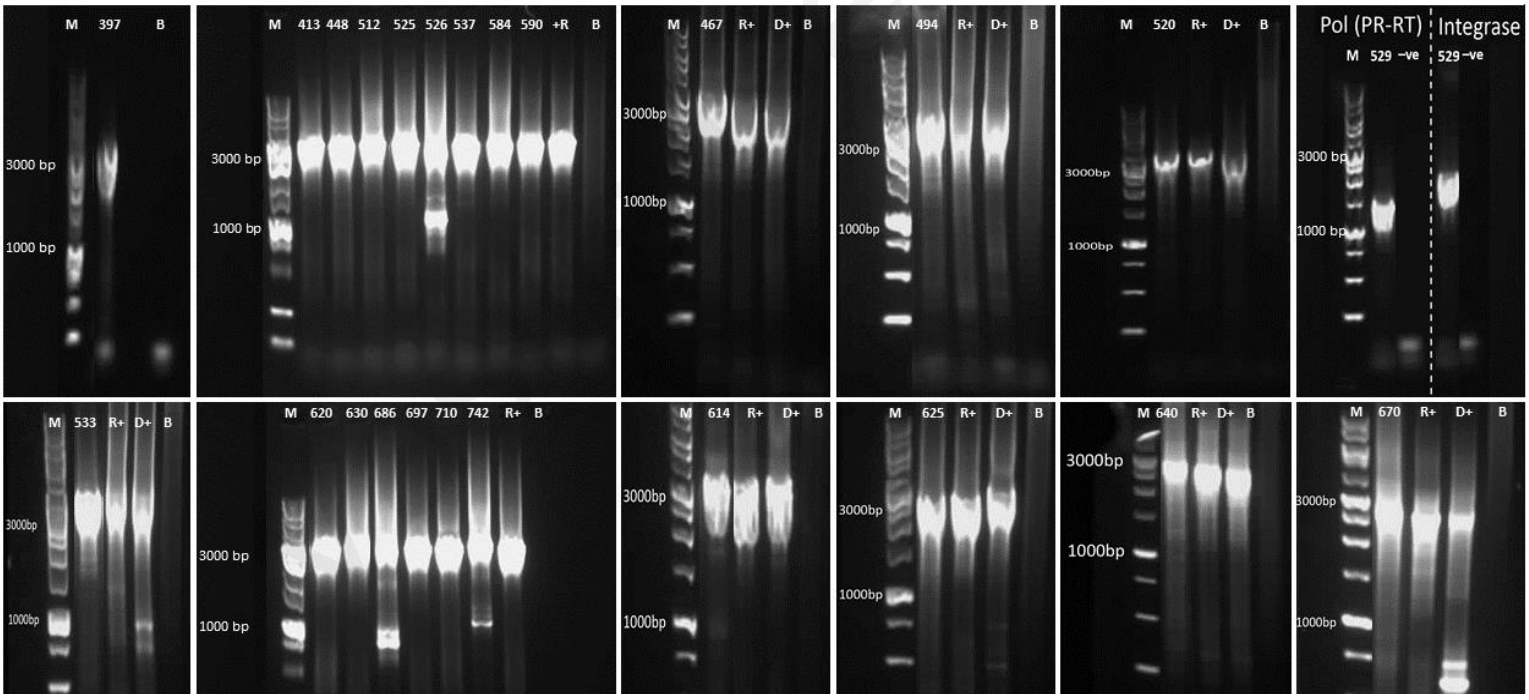
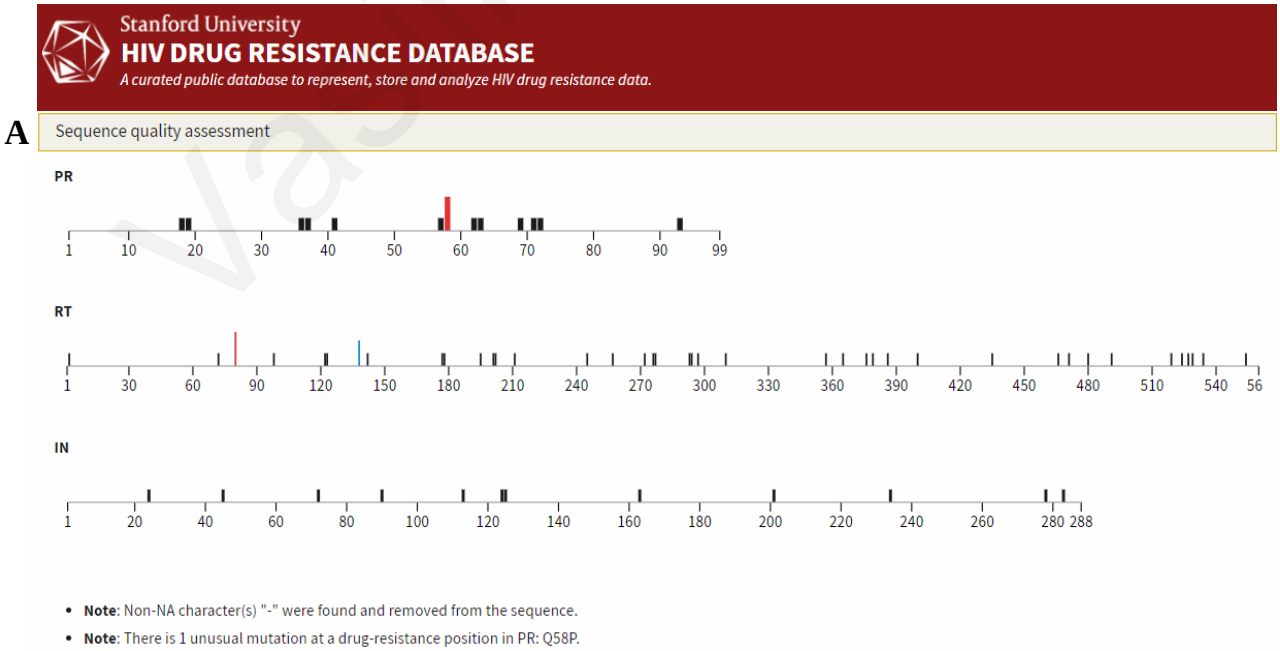


Figure 17: Gel electrophoresis of the pol region obtained from the samples of interest. The samples are represented in separate gels because they were processed in different time peridots. M: DNA ladder 1Kb; R+: Positive RNA control; D+: Positive DNA control; B: Blank control. The positive controls refer to known positive RNA and DNA samples used to validate the applied methodologies. The Blank control contained all the reagents except the HIV-1 samples, and it was used for the validation in terms of contamination.

3.3 Dug Resistance Analysis

After the quality analysis, the sequences were subsequently imported into the Stanford HIV Drug Resistance Database (Tang, Liu et al. 2012) for the investigation of possible drug resistance mutations. The HIV-1 Drug Resistance database has the ability to identify more than 200 mutation patterns (Bennett, Camacho et al. 2009) but also provides in depth results since it recognizes unusual mutations as well as mutations caused by the APOBEC enzymes. Stanford HIV-1 Drug Resistance Database categorizes antiretroviral therapy resistance into four groups (1) resistance to the nucleoside reverse transcriptase inhibitors (NRTIs), (2) resistance to the non-nucleoside reverse transcriptase inhibitors (NNRTIs), (3) resistance to the integrase inhibitors (INSTIs), (4) resistance to the protease inhibitors (PIs). The drug resistance analyses that was performed for the study subjects of this research showed that none of the patients had any mutations that cause high resistance to the antiretroviral therapy. However, several patients are characterized by mutations that enable the virus to have low or potential low resistance on specific drugs. In particular, CY620, CY670 and CY710 have the polymorphic mutation A71V/T. This mutation is a PI-selected accessory mutation that increases the replication of virus in combination with other PI-resistance mutations. Furthermore, the same patients have the common polymorphic NNRTI accessory mutation E138A, that is weakly selected in patients receiving etravirine (ETR) and rilpivirine (RPV) and it cause the reduction of ETR and RPV susceptibility by ~2-fold (Figure 18).



B Drug resistance interpretation: PR — HIVDB 9.0 (2021-02-22)

PI Major Resistance Mutations:	None
PI Accessory Resistance Mutations:	None
Other Mutations:	Q18QK, L19I, M36V, N37T, R41K, R57K, Q58QP, I62V, L63S, H69N, A71V, I72T, I93L

Protease Inhibitors	
atazanavir/r (ATV/r)	Susceptible
darunavir/r (DRV/r)	Susceptible
lopinavir/r (LPV/r)	Susceptible

PR comments

Other

- **A71V/T** are polymorphic, PI-selected accessory mutations that increase the replication of viruses with other PI-resistance mutations.

C Drug resistance interpretation: RT — HIVDB 9.0 (2021-02-22)

NRTI Resistance Mutations:	None
NNRTI Resistance Mutations:	**E138A**
Other Mutations:	I2IL, R72RG, L80LF, A98S, K122P, D123E, I142V, D177E, I178V, I195L, K201R, I202V, R211K, V245E, I257V, A272P, V276VI, K277KR, I293V, P294A, E297K, L310I, M357T, V365I, A376S, S379G, T386I, A400T, V435M, V466A, D471E, Q480H, L491S, S519N, Q524K, K527E, E529D, A534S, A554S

Nucleoside Reverse Transcriptase Inhibitors	
abacavir (ABC)	Susceptible
zidovudine (AZT)	Susceptible
emtricitabine (FTC)	Susceptible
lamivudine (3TC)	Susceptible
tenofovir (TDF)	Susceptible

Non-nucleoside Reverse Transcriptase Inhibitors	
doravirine (DOR)	Susceptible
efavirenz (EFV)	Susceptible
etravirine (ETR)	Potential Low-Level Resistance
nevirapine (NVP)	Susceptible
rilpivirine (RPV)	Low-Level Resistance

RT comments

NNRTI

- **E138A** is a common polymorphic accessory mutation weakly selected in patients receiving ETR and RPV. It reduces ETR and RPV susceptibility ~2-fold. It has a weight of 1.5 in the Tibotec ETR genotypic susceptibility score.

Dosage Considerations

- This virus is predicted to have low-level reduced susceptibility to **RPV**. The use of the combination of CAB/**RPV** should be considered to be relatively contraindicated.

D Drug resistance interpretation: IN — HIVDB 9.0 (2021-02-22)

IN Major Resistance Mutations:	None
IN Accessory Resistance Mutations:	None
Other Mutations:	S24SG, L45LI, I72V, P90S, I113V, T124N, T125V, G163E, V201I, L234I, D278N, S283G

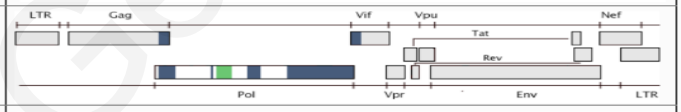
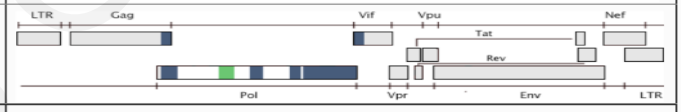
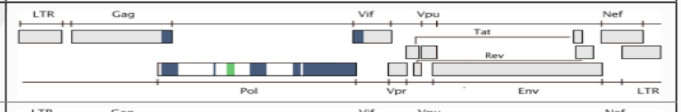
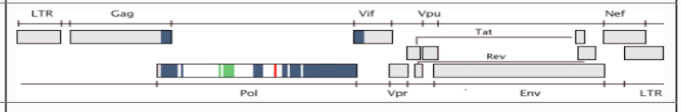
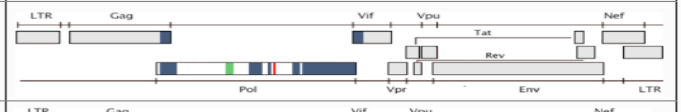
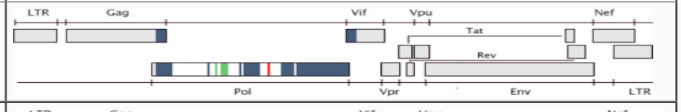
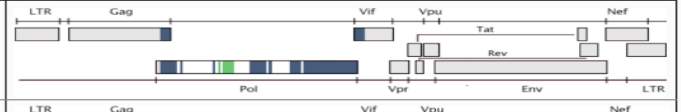
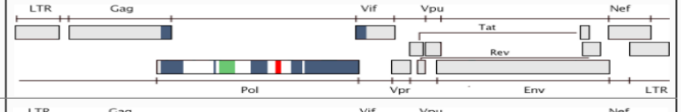
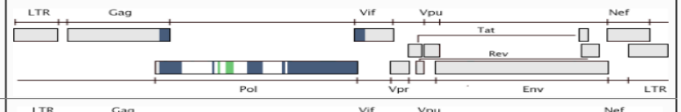
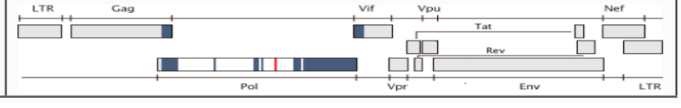
Integrase Strand Transfer Inhibitors	
bictegravir (BIC)	Susceptible
cabotegravir (CAB)	Susceptible
dolutegravir (DTG)	Susceptible
elvitegravir (EVG)	Susceptible
raltegravir (RAL)	Susceptible

Figure 18: *Drug resistance analysis of CY620 through the Stanford HIV-1 Drug Resistance* database. The figure illustrates an example of the analysis produced by the Stanford HIV-1 Drug Resistance database. (A) The first section indicates the sequence quality assessment that recognizes the mutations that exist in the PR, RT and IN genes. The black lines indicate the common/polymorphic mutations, the blue lines indicate the accessory mutations and the red lines indicate the unusual mutations. The (B), (C), and (D) sections indicate the drug resistance interpretation of the PR, RT and IN genes accordingly. In this section the mutations are categorized either as resistance mutations or other mutations. If a mutation enables the virus to have potential low-level, low-level, intermediate level or high-level resistance then the mutation is described further in the comment section at the end of each gene drug resistance interpretation. In this example the CY620 is characterized by the A71V/T PI-selected accessory mutation and the common polymorphic NNRTI accessory mutation E138A.

3.4 Subtype Characterization

After the drug resistance analysis, the subtype identification of the pol region (2253-5250 in the HXB2 genome) was followed using the REGA HIV-1 Subtyping Tool version 3 (Pineda-Peña, Faria et al. 2013). The REGA HIV-1 Subtyping Tool uses phylogenetic methods to detect the subtype of a particular sequence and bootscan methods for the recombination analysis. The initial result of the REGA HIV-1 Subtyping Tool includes the name of the samples, the length, the assignment of the sequence, and the genome presented in different colors, where each color symbolizes a different subtype or recombinant strain (Table 12 – Table 15). In addition, REGA HIV-1 Subtyping Tool provides the ability to evaluate the report since it delivers the data and the reference dataset used to extract the results.

Table 12: Cluster 9 "Rec. of 02_AG, G, B" pol Subtype Characterization Based on REGA Subtyping Tool

Name	Length	Assignment	Genome
CY467	2998	Recombinant of G, A1, B	
CY494	2998	Recombinant of 02_AG, G	
CY520	2998	Recombinant of 02_AG, G	
CY533	2998	Recombinant of 02_AG, G	
CY614	2999	Recombinant of 02_AG, G, A1	
CY622	2998	Recombinant of 02_AG, G, A1	
CY630	2998	Recombinant of 02_AG, G, A1	
CY640	2998	Recombinant of 02_AG, G	
CY670	2998	Recombinant of 02_AG, G, A1	
CY686	2998	Recombinant of 02_AG, G	
CY742	2998	Recombinant of 02_AG, A1	

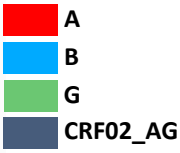


Table 13: Cluster 4 "Rec. of B, A1, G" pol Subtype Characterization Based on REGA Subtyping Tool

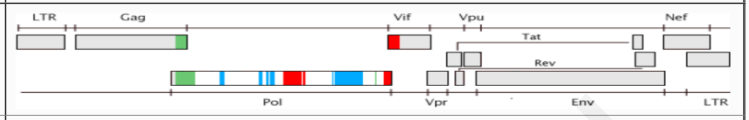
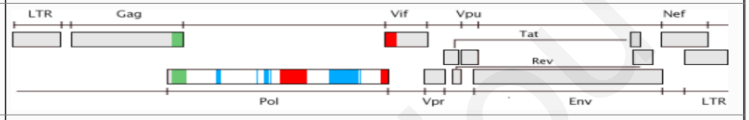
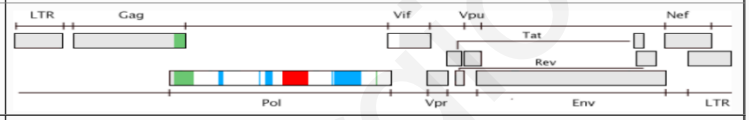
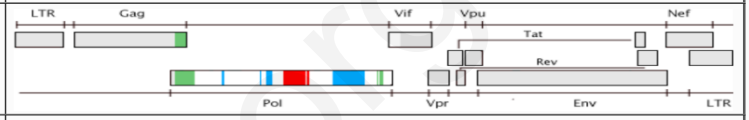
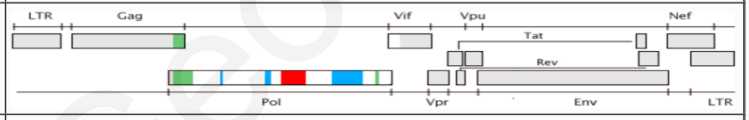
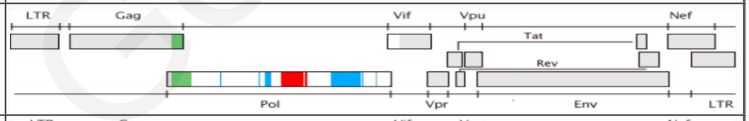
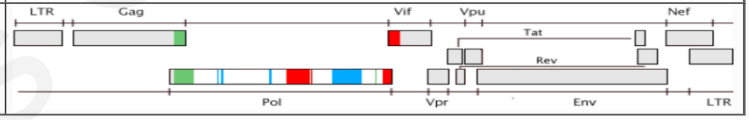
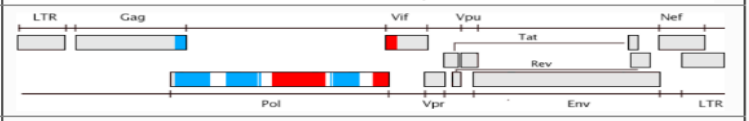
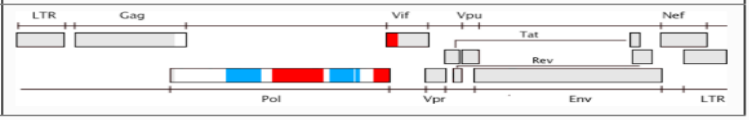
Name	Length	Assignment	Genome
CY448	2998	Recombinant of B, A1, G	
CY512	2998	Recombinant of B, A1, G	
CY525	2998	Recombinant of B, A1, G	
CY526	2998	Recombinant of B, A1, G	
CY529	2998	Recombinant of B, A1, G	
CY537	2998	Recombinant of B, A1, G	
CY625	2998	Recombinant of B, A1, G	

Table 14: Cluster 5 "Rec. of B, A1" pol Subtype Characterization Based on REGA Subtyping Tool

Name	Length	Assignment	Genome
CY397	2998	Recombinant of B, A1	
CY413	2998	Recombinant of B, A1	
CY590	2998	Recombinant of B, A1	

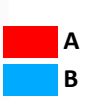
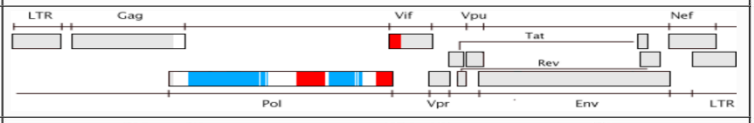
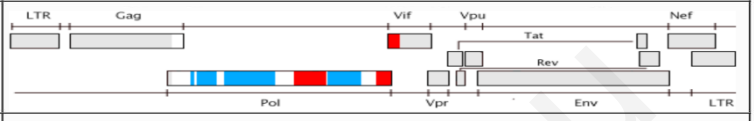
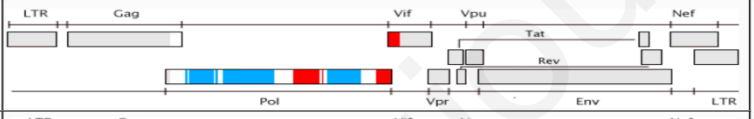
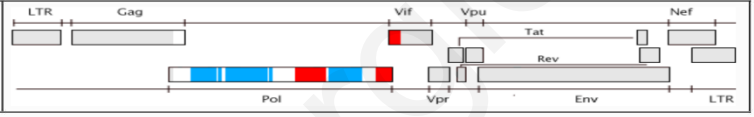


Table 15: Cluster 16 pol "Rec. of B, A1" Subtype Characterization Based on REGA Subtyping Tool

Name	Length	Assignment	Genome
CY584	2998	Recombinant of B, A1	
CY620	2998	Recombinant of B, A1	
CY697	2998	Recombinant of B, A1	
CY710	2998	Recombinant of B, A1	

3.5 Phylogenetic Analysis

Next monthly phylogenetic analysis of the pol region (2253-5250 in the HXB2 genome) was performed as part of the near real-time surveillance of the HIV-1 transmission dynamics in Cyprus (https://www.kostrikislab.com/ongoing-projects/). This analysis was performed every month from April 2019 until September 2021, and each month the new samples that were obtain through the Grigorios HIV Clinic of Larnaca General Hospital were included. Reaching the final September 2021 monthly report, all the samples received by the BMV lab from March 2017 until October of 2021 were included (Figure 19). The phylogenetic analysis was executed as described at the annex 2.10. The results of the subtype characterization and phylogenetic analysis of the pol region (2253-5250 in the HXB2 genome) indicate the existence of 31 samples that formed four recombinant clusters. The first recombinant cluster, cluster 4 "Rec. of B, A1, G" was identified on the April 2019 phylogenetic report, followed by cluster 5 "Rec. of B, A1" on June 2019, and cluster 9 "Rec. of 02_AG, G, B" along with cluster 16 "Rec. of B, A1" on August 2019. From April 2019 until September 2021 the cluster 4 "Rec. of B, A1, G", and 5, 16 "Rec. of B, A1" indicated a minor growth and the cluster 9 "Rec. of 02_AG, G, B" an active growth. Specifically, cluster 4 "Rec. of B, A1, G", cluster 5 "Rec. of B, A1" and cluster 16 "Rec. of B, A1" were increased by one sample, and cluster 9 "Rec. of 02_AG, G, B" was increased by 12 samples.

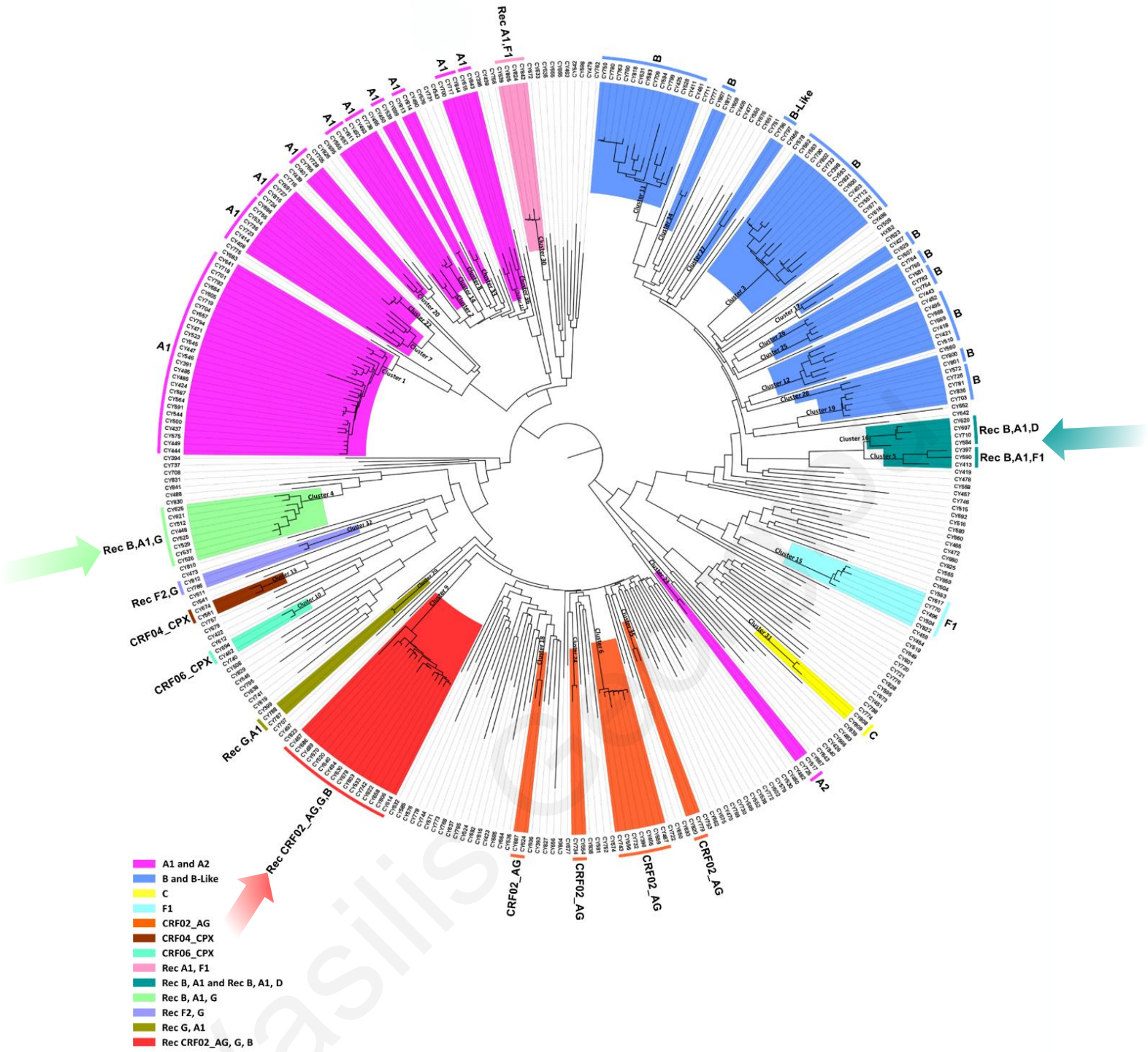


Figure 19: *September 2021 Monthly Phylogenetic analysis* (https://www.kostrikislab.com/ongoing-projects/). The figure illustrates the September 2021 monthly phylogenetic report of the BMV laboratory of the University of Cyprus. The different colors of the branches of the maximum likelihood tree represent the distinct transmission clusters that circulate in Cyprus. The colored arrows illustrate the recombinant transmission clusters 4, 5, 9 and 16 that contain the study subjects of this research. All the samples of the phylogenetic analysis are part of the 2017-2021 cohort of the BMV laboratory.

Those evidence led to the hypothesis that those recombinant clusters are potential novel CRFs and URFs in Cyprus. Pursuing this hypothesis, the amplification and sequencing of the gag (790-2292 in the HXB2 genome) (annex 3.6) and env region (5041-8795 in the HXB2 genome) (annex 3.7) was perform with the aim to obtain the near full-length HIV-1 genome sequences of the potential recombinant samples.

3.6 Amplification and Sequencing of the *gag* region

For the amplification and sequencing of the gag region (790-2292 in the HXB2 genome) the same methodologies as the amplification and sequencing of the pol region (2253-5250 in the HXB2 genome) (annex 3.2) were applied, using the PCR and sequencing primers shown in the Table 1 and the primary RT-PCR and secondary (NESTED) PCR thermocycling conditions as indicated in the Table 3 and Table 6 accordingly. Based on the amplification primers of the secondary (NESTED) PCR (Table 1) the expected amplified fragment of the HIV-1 was 1502 bp and as such on the gel images is represented at approximately 1500 bp according to the 1Kb DNA ladder. The positive samples were then sent for Sanger sequencing to Macrogen (Macrogen-Europe). Subsequently, the quality of the sequencing results were evaluated using Geneious software (Kearse, Moir et al. 2012) (annex 2.9).

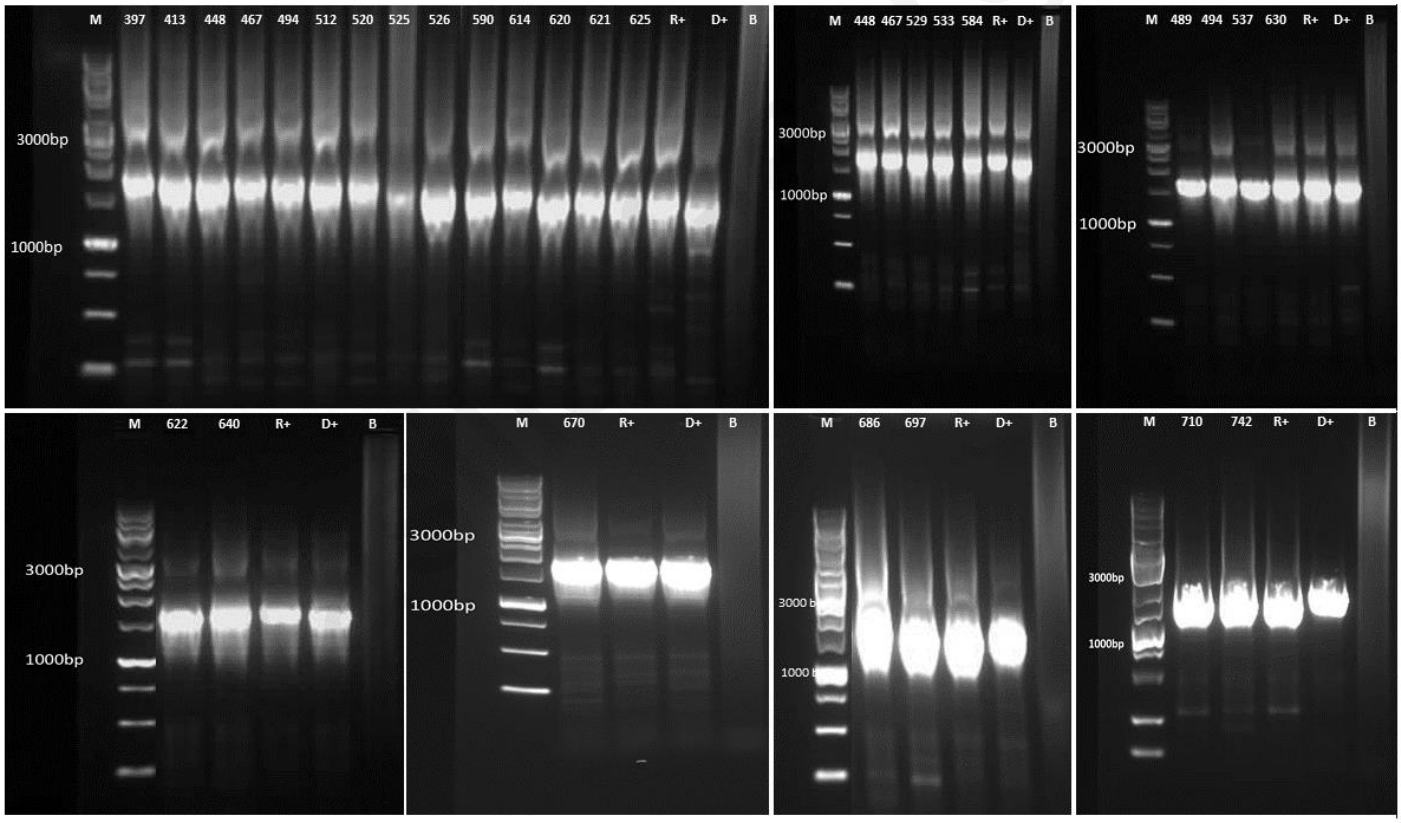


Figure 20: Gel electrophoresis of the gag region obtained from the samples of interest. The samples are represented in separate gels because they were processed in different time peridots. M: DNA ladder 1Kb; R+: Positive RNA control; D+: Positive DNA control; B: Blank control. The positive controls refer to known positive RNA and DNA samples used to validate the applied methodologies. The Blank control contained all the reagents except the HIV-1 samples, and it was used for the validation in terms of contamination.

3.7 Amplification and Sequencing of the env region

For the amplification and sequencing of the env region (5041-8795 in the HXB2 genome) the same methodologies as the amplification and sequencing of the pol and gag regions (annex 3.2 and annex 3.6) were applied, using the PCR and sequencing primers according to the Table 1 and the Primary RT-PCR and secondary (NESTED) PCR thermocycling conditions as indicated in the Table 3 and Table 7 accordingly. For the env region at the secondary (NESTED) PCR a different DNA *pol*ymerize enzyme (Invitrogen™ Platinum™ SuperFi™ II) was used that has the ability to amplify larger genome fragments and as such the thermocycling conditions of the secondary (NESTED) PCR were modified according to the manufacturer's instructions (Table 7). Based on the amplification primers of the secondary (NESTED) PCR (Table 1) the expected amplified fragment of the HIV-1 was 5404 bp and as such on the gel images is represented at approximately 5000 bp according to the 1Kb DNA ladder. The positive samples were then sent for Sanger sequencing to Macrogen (Macrogen-Europe). Subsequently, the quality of the sequencing results were evaluated using Geneious software (Kearse, Moir et al. 2012) (annex 2.9). At the end of the qualitative analysis gag, pol and env sequences of each sample were aligned through Geneious software leading to the generation of the near full-length HIV-1 genome sequence of the samples of interest.

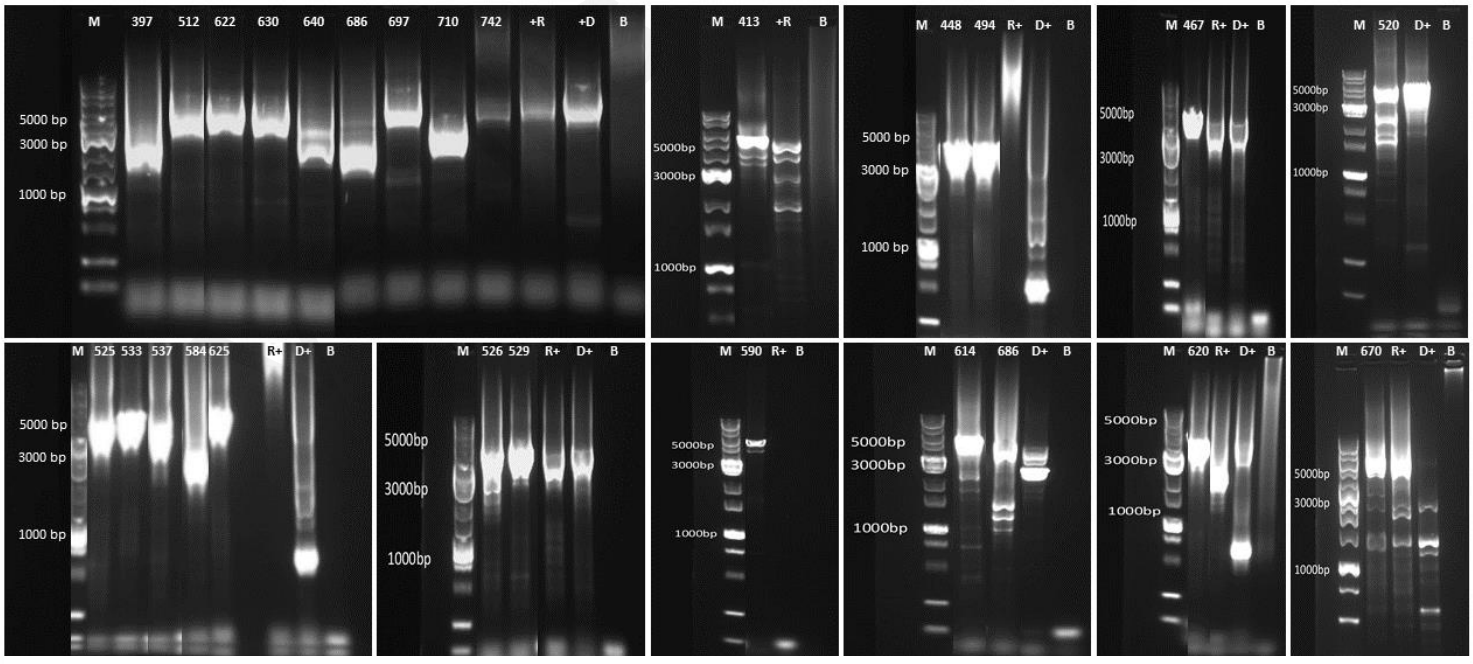


Figure 21: Gel electrophoresis of the env region obtained from the samples of interest. The samples are represented in separate gels because they were processed in different time peridots. M: DNA ladder 1Kb; R+: Positive RNA control; D+: Positive DNA control; B: Blank control. The positive controls refer to known positive RNA and DNA samples used to validate the applied methodologies. The Blank control contained all the reagents except the HIV-1 samples, and it was used for the validation in terms of contamination.

3.8 Recombination Analysis of the Mosaic Genome of the Recombinant Clusters

The successful sequencing of the near-full genome of the 24 samples that belonged to the recombinant clusters (Figure 19) was followed by the recombination analysis in order to investigate for potential novel CRFs and URFs in Cyprus. For this purpose, several bootscan, similarity and phylogenetic approaches were used in order to crosscheck every possibility and as such to obtain highly reliable results. Through maximum likelihood trees the samples were compared with well-established pure subtypes and recombinants of the HIV-1 M group by using the RIP 2017 and RIP 2020 alignment obtained from the Los Alamos HIV sequence Database (www.hiv.lanl.gov/) (Compendium 2018). The maximum likelihood trees were created by using IQTREE v.2.1.2 software (Nguyen, Schmidt et al. 2015, Trifinopoulos, Nguyen et al. 2016), and MEGA-X software (Kumar, Stecher et al. 2018) as described in annex 2.10. Subsequently, to characterize the mosaic genome, the near full-length HIV-1 genome sequences were analyzed through REGA HIV-1 Subtyping Tool (Pineda-Peña, Faria et al. 2013), Simplot program vesion 3.5.1 (Lole, Bollinger et al. 1999), jumping profile hidden Markov model (jpHMM) (Schultz, Bulla et al. 2012) and the Recombinant Identification Program (RIP) tool (SIEPEL, HALPERN et al. 1995) of the Los Alamos HIV sequence Database (www.hiv.lanl.gov/) (Compendium 2018) as described in annex 2.11.To verify the results of the aforementioned approaches, neighbor joining phylogenetic trees were created for each fragment of the genome of each recombinant sample separately using MEGA-X software (Kumar, Stecher et al. 2018) as described in annex 2.10 and 2.11.

For each recombinant sample separately, all the aforementioned approaches revealed identical results. As such, in order to avoid repetitive information, only the maximum likelihood trees of the clusters, along with the similarity plots, bootscan analysis and the neighbor joining phylogenetic trees of the inter-breakpoint segments of the representative samples of each cluster are illustrated in this study. The recombination analysis of the remaining samples of each cluster could be found in the supplementary data .

The overall outcome of the recombination analysis was the identification of the novel CRF91_cpx along with a URF consisted of CRF91_cpx, B (cluster 9) (annex 3.8.1), and evidence

of the existence of three other putative CRFs. One CRF consisted of CRF56_cpx, G (cluster 4) (annex 3.8.2) and the two others consisted of subtype A1, B (cluster 5 and 16) (annex 3.8.3).

3.8.1 Recombination Analysis of the Cluster9 (CRF91_cpx)

The initial analysis was the construction of a maximum likelihood tree to identify the uniqueness of the recombinant samples of the cluster 9 "Rec. of 02_AG, G, B". This analysis was performed using the RIP 2020 reference dataset obtained through the Los Alamos HIV sequence Database (www.hiv.lanl.gov/) (Compendium 2018). The RIP 2020 reference dataset contains 198 near full-length HIV-1 sequences of all pure subtypes along with 118 CRFs and the N, O and P outgroups. As indicated on the Figure 22, the samples of interest cluster together with branch support value of 100%. The most genetically similar reference sequences with the samples of interest was a cluster of CRF06_cpx and CRF06_cpx recombinant sequences with the common ancestor to be supported by 65% branch value.

Subsequently, for further validation of the uniqueness of the recombinant samples of the cluster 9 "Rec. of 02_AG, G, B", blast analysis was performed through the NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and the Los Alamos HIV sequence database (https://www.hiv.lanl.gov/content/sequence/BASIC_BLAST/basic_blast.html) in order to find possible identical sequences. The results of the blast analysis indicate an 89.44% similarity to a recombinant strain CRF02_AG / CRF06_cpx (Accession number: AJ508595) and a similarity equal or less than 89% with CRF02_AG recombinant strains. The overall outcome of the aforementioned analysis was that the recombinant samples of cluster 9 "Rec. of 02_AG, G, B" didn't present high genetic similarity to any of the known HIV-1 pure subtypes or CRFs.

Next, several approaches were conducted with the aim to characterize the mosaic genome of the near full-length HIV-1 genome sequences of cluster 9 "Rec. of 02_AG, G, B". Pursuing this purpose all the approaches that were performed (annex 2.11) indicated the same results. However to avoid repetitive information, only the results of the Simplot program vesion 3.5.1 (Lole, Bollinger et al. 1999) and Neighbor joining phylogenetic trees using MEGA-X software (Kumar, Stecher et al. 2018) of the representative sample are illustrated as part of this research. The results of the rest of the samples of cluster 9 "Rec. of 02_AG, G, B" can be found in the supplementary data (Figure S1- S9).

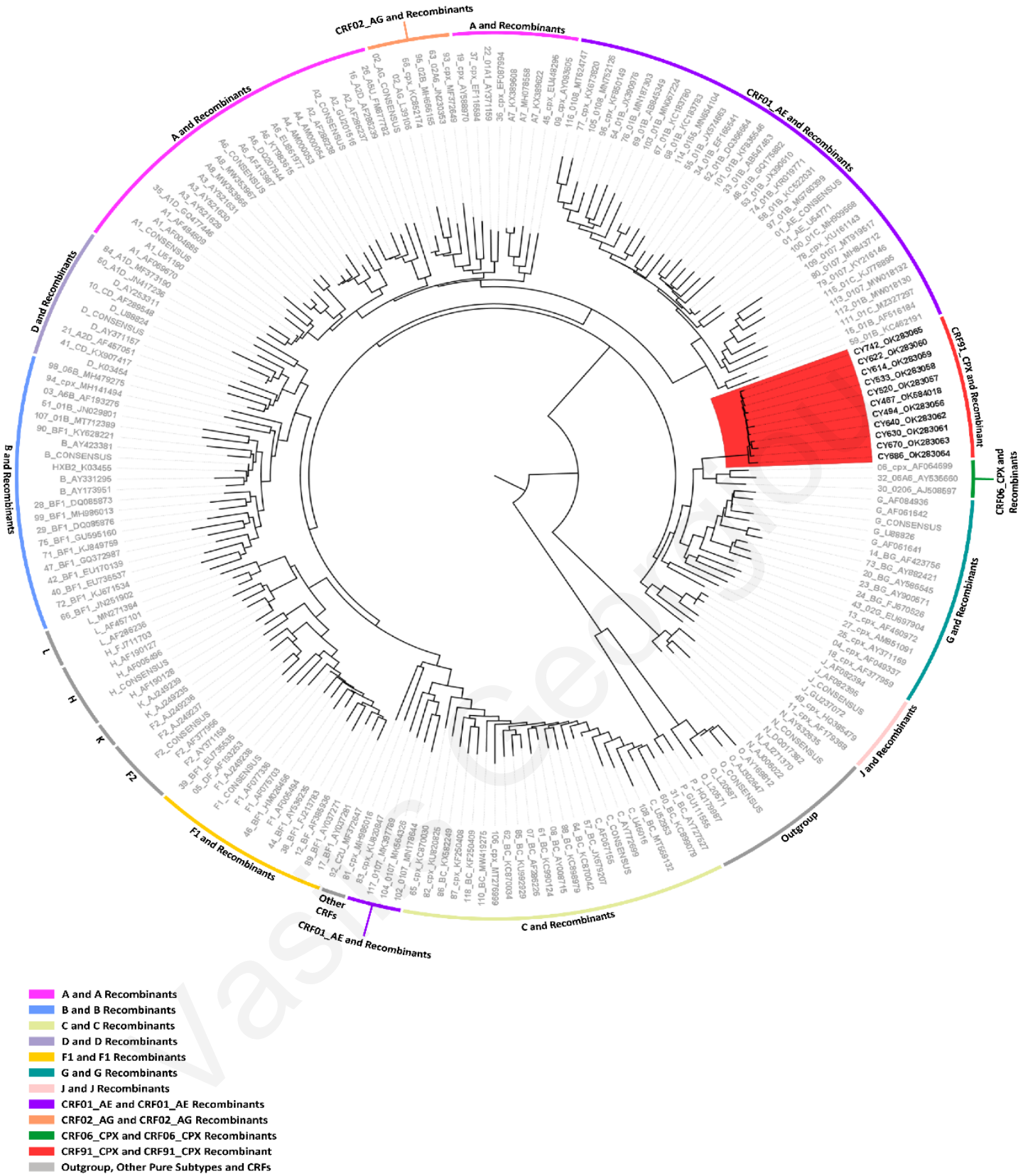


Figure 22: *Maximum likelihood tree analysis of the NFLG sequences of cluster 9 with standard references including the M group pure subtypes and CRFs.* The branch indicated with the red color represents the samples of cluster 9. The outer circular colors indicate the different subtypes and recombinant clusters of the reference dataset used for the creation of the ML tree (RIP 2020 alignment). The cluster that has the closest genetic similarity with the cluster 9 is indicated with the green color, which consist of CRF06_cpx and CRF06_cpx recombinant samples. The gray color represents several clusters of pure subtypes that have low prevalence in Cyprus, the outgroups and complex CRFs.

The initial investigation of the mosaic genome of the near full-length HIV-1 sequences of cluster 9 "Rec. of 02_AG, G, B" revealed an identical mosaic genomic patterns that was composed of CRF02_AG, subtypes G and subtypes J. Specifically, the consensus mosaic genome pattern was composed from seven fragments, beginning from the 5' end with CRF02_AG from nucleotide position 790 to 3034 with average branch support value of 100% ; from 3035 to 3082, recombination range between the subtypes CRF02_AG and subtype G; from 3082 to 3410, was characterized by subtype G-like with average branch support value of 74.4%; from 3411 to 3484, recombination range between the subtypes G and CRF02_AG; from 3485 to 5421, was characterized by CRF02_AG with average branch support value of 100%; from 5422 to 6622, was characterized by subtype G with average branch support value of 100%; from 6623 to 7553 wasn't characterized and as such is termed as an unidentified region; from 7554 to 8433 was characterized by subtype G with average branch support value of 99.8%; and from 8434 to 8686, was characterized by subtype J-like with average branch support value of 76.1% (Figure 24). Based on Blast analysis that was conducted for each fragment of the genome separately the last fragment (8434-8686) indicated high similarity score with CRF06_cxp samples. Thus, a second phylogenetic analysis was performed for the last fragment (from 8434 to 8686 bp) that illustrated a 100% average brunch support value with the CRF06_cpx reference sequences. It is important to note that CRF06_cpx at the nucleotide position from 8434 to 8686 is also characterized by the subtype J and as such the last fragment was determine as J-like subtype (Figure 25).

The investigation of the mosaic genome of the near full-length HIV-1 sequences of cluster 9 "Rec. of 02_AG, G, B" led also to the identification of a Unique Recombinant Form that was composed of CRF91_cpx and the subtype B (sample CY467). This result comes with an agreement with the preliminary phylogenetic analysis of the pol region (2253-5250 in the HXB2 genome) (Figure 19) that presented a diverge of CY467 from the rest of the samples of the cluster 9 "Rec. of 02_AG, G, B". In particular, the genomic mosaic pattern of CY467 was characterized by CRF91_cpx with 100% branch support from nucleotide position 790 to 3803; from 3804 to 4280 was characterized by subtype B with branch support 81%; and from 4281 to 8795 was characterized by CRF91_cpx with branch support 100%. The mosaic genomic pattern of the sample CY467 (URF of CRF91_cpx, B) is represented on Figure 23 below.

Following the recombination analysis, the sequences along with the mosaic genome and breakpoints of the samples were sent to the Los Alamos HIV sequence Database (www.hiv.lanl.gov/) (Compendium 2018). After an extensive analysis by the Los Alamos HIV sequence Database experts, their results came with an agreement with the aforementioned results leading to the identification of the novel CRF91_cpx, named according to the nomenclature of HIV-1 (Robertson, Anderson et al. 2000).

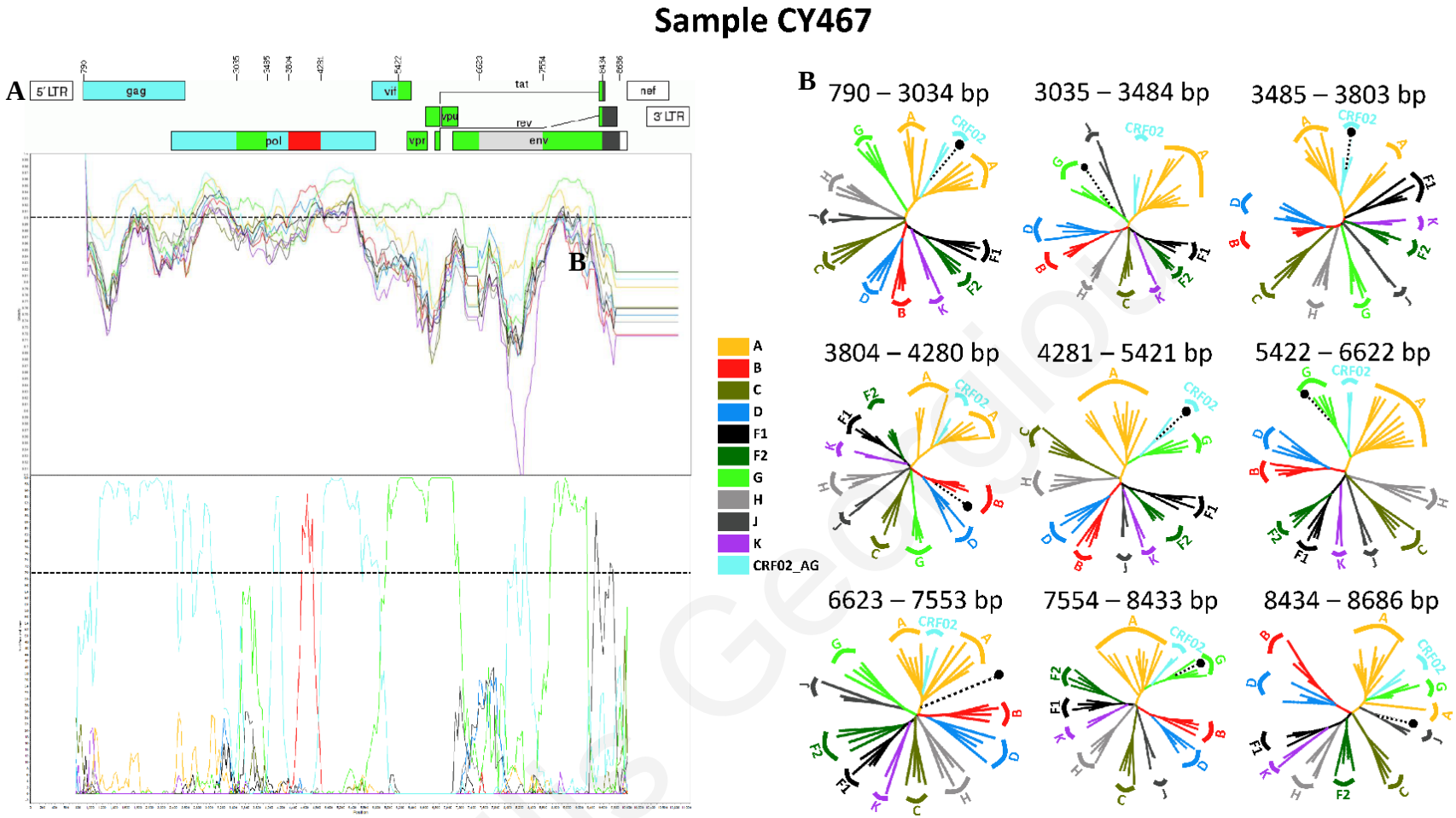


Figure 23: *Recombination analysis of the Unique Recombinant Form CRF91_cpx, B (Sample CY467).* The left part of the figure (A) represents the mosaic structure of the near full-length genome of the CY467 isolate in comparison to reference strains of all HIV-1 pure subtypes of the M group along with the CRF02_AG recombinant strain. The upper diagram shows the gene regions and the recombination breakpoints of the CY467 isolate as defined by bioinformatic analysis. The illustration was created according to HXB2 numbering using the Recombinant Drawing tool available on the Los Alamos HIV sequence Database website *(https://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html).* The middle panel demonstrate the similarity plot diagram, and the bottom panel shows the bootscan analysis, created by Simplot program vesion 3.5.1 (Lole, Bollinger et al. 1999). The y-axis in the similarity plot indicates the percent identity of the query sequence to a set of reference sequences and the dotted line shows the 90% (significant) similarity; in the bootscan diagram, the y-axis indicates the bootstrap value, and the dotted line indicates 70% (significant) bootstrap value. The x-axis shows the nucleotide position of the HXB2 genome. The right part of the Figure (B) indicates the phylogenetic analysis of the inter-breakpoint segments comprises the CY467 strain as defined by the similarity plot and the bootscan analysis. Analysis was performed using the NJ method with the Kimura's two-parameter parameter distance estimation method and bootstrap analysis of 1000 replicates. All reference HIV-1 pure subtypes and recombinants of the M group that were used to construct the trees are denoted with different colors. All the inter-breakpoint segments results are supported by bootstrap value ≥70%; Above of each NJ tree there is the loci of genome segments based on the HXB2 numbering. The reference sequences of the pure subtype were obtained by the RIP 2017 alignment and the CRF02_AG reference samples were obtained through blast analysis.

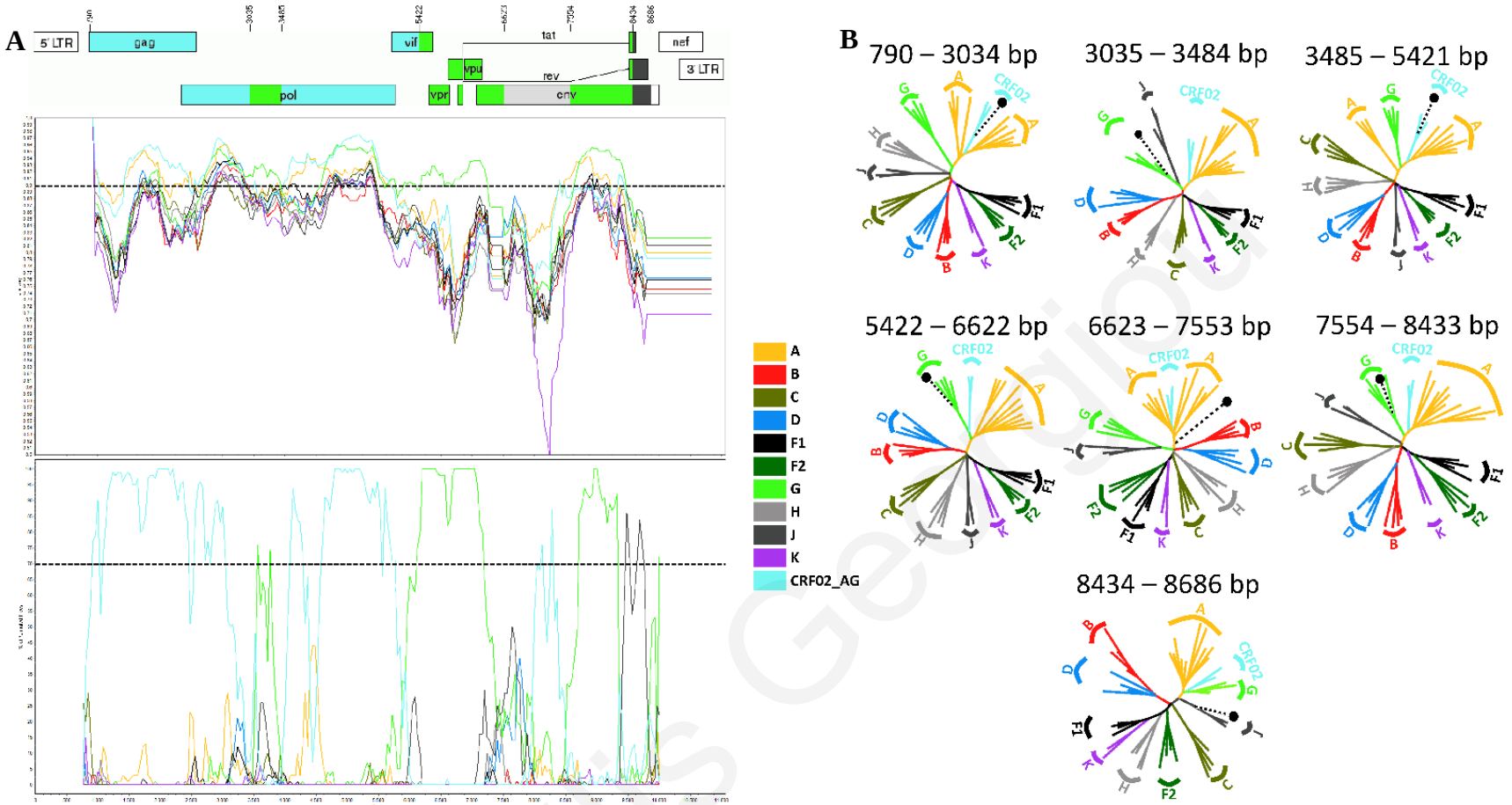


Figure 24: *Recombination analysis of the CRF91_cpx representative Sample CY494.* The left part of the figure (A) represents the mosaic structure of the near full-length genome of the CY494 isolate in comparison to reference strains of all HIV-1 pure subtypes of the M group along with the CRF02_AG recombinant strain. The upper diagram shows the gene regions and the recombination breakpoints of the CY494 isolate as defined by bioinformatic analysis. The illustration was created according to HXB2 numbering using the Recombinant Drawing tool available on the Los Alamos HIV sequence Database website *(https://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html).* The middle panel demonstrate the similarity plot diagram, and the bottom panel shows the bootscan analysis, created by Simplot program vesion 3.5.1 (Lole, Bollinger et al. 1999). The y-axis in the similarity plot indicates the percent identity of the query sequence to a set of reference sequences and the dotted line shows the 90% (significant) similarity; in the bootscan diagram, the y-axis indicates the bootstrap value, and the dotted line indicates 70% (significant) bootstrap value. The x-axis shows the nucleotide position of the HXB2 genome. The right part of the Figure (B) indicates the phylogenetic analysis of the inter-breakpoint segments comprises the CY494 strain as defined by the similarity plot and the bootscan analysis. Analysis was performed using the NJ method with the Kimura's two-parameter parameter distance estimation method and bootstrap analysis of 1000 replicates. All reference HIV-1 pure subtypes and recombinants of the M group that were used to construct the trees are denoted with different colors. All the inter-breakpoint segments results are supported by bootstrap value ≥70%; Above of each NJ tree there is the loci of genome segments based on the HXB2 numbering. The reference sequences of the pure subtype were obtained by the RIP 2017 alignment and the CRF02_AG reference samples were obtained through blast analysis.

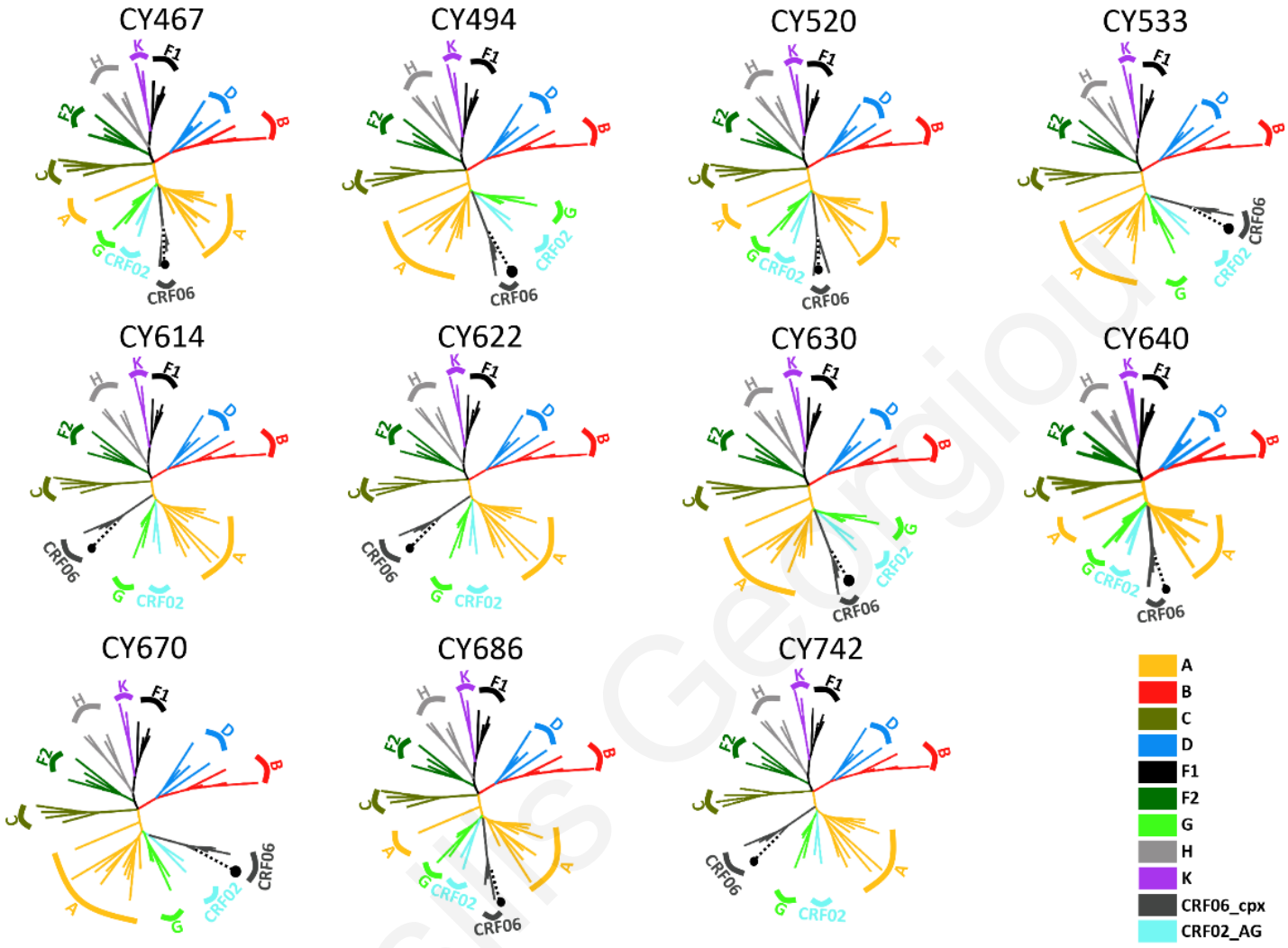


Figure 25: *Further investigation for the subtype identification of the inter-breakpoint segment 8434-8686 of the CRF91_cpx.* The figure illustrates the phylogenetic analysis of the inter-breakpoint segments 8434-8686 of all the CRF91_cpx samples and the URF of CRF91_cpx, B (CY467). The analysis was performed using the NJ method with the Kimura's two-parameter parameter distance estimation method and bootstrap analysis of 1000 replicates. All reference HIV-1 pure subtypes and recombinants of the M group that were used to construct the trees are denoted with different colors. For this analysis the pure subtype J sequences was exchange with CRF06_cpx sequences obtained through blast analysis. All the inter-breakpoint segment results for all the samples are supported by bootstrap value of 100%; Above of each NJ tree there is the sample Lab ID. The reference sequences of the pure subtype were obtained by the RIP 2017 alignment and the CRF02_AG, CRF06_cpx reference samples were obtained through blast analysis.

After the identification of the CRF91_cpx an analysis was performed with the aim to identify near full-length genome sequences or partial sequences of CRF91_cpx that circulate around the world. Following this goal near full-length genome and sub-section Blast analysis was performed along with similarity-based approaches (annex 2.11). The sub-section Blast analysis, included all the inter-breakpoint segments of the CRF91_cpx along with the full pol gene and the env variable region 1 - variable region 3. The pol and env segments were selected based on the vast number of pol region sequences that have been uploaded into GenBank due to interest in drug resistance mutations and the vast number of env sequences due to the interest in vaccine design and HIV-1 evolution. This investigation revealed two partial pol sequences of Nigerian patients (GenBank accession numbers: MH654941; MH654876) that indicated high similarity score with the CRF91_cpx samples.

Based on the high similarity score of those two sequences, similarity plot, bootscan and phylogenetic analysis was then conducted. Bootscan analysis and similarity plots were created using Simplot program version 3.5.1 (Lole, Bollinger et al. 1999) (annex 2.11). The phylogenetic analysis was based on maximum likelihood estimation method using IQTREE v.2.1.2 software (Nguyen, Schmidt et al. 2015, Trifinopoulos, Nguyen et al. 2016), as described in annex 2.10 (data not shown). The results revealed that the two partial pol region sequences of the Nigerian patients (GenBank accession numbers: MH654941; MH654876) share the same mosaic structure with the CRF91_cpx samples (Figure 26 and Figure 27).

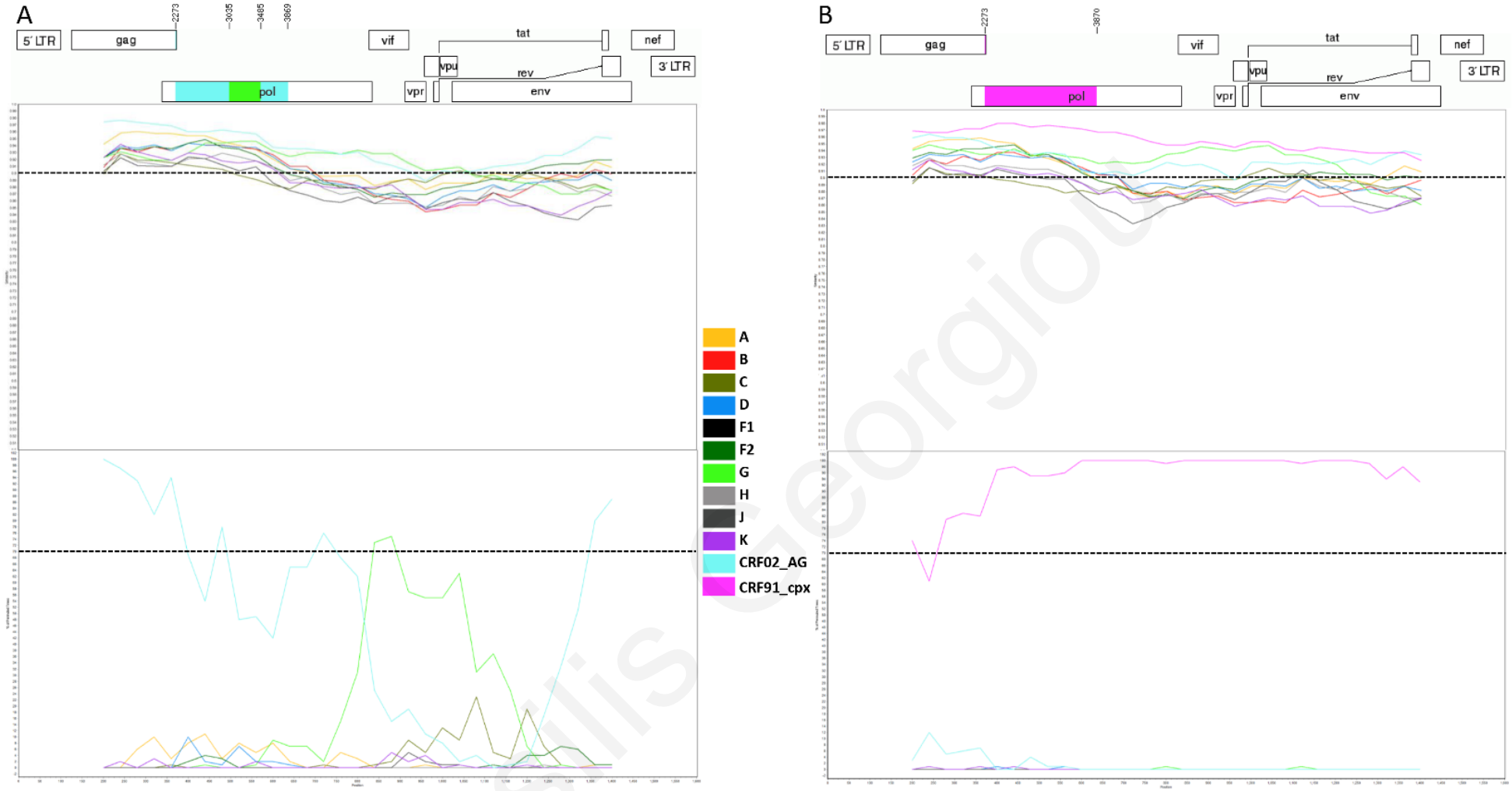


Figure 26: *Identification of CRF91_cpx* pol *partial genome Circulating in Nigeria (GenBank accession number: MH654941)*. (A) The upper diagram shows the identical recombination breakpoints of the sample MH654941 isolate with the CRF91_cpx mosaic pattern as defined by bioinformatic analysis using as reference dataset the pure subtypes of the HIV-1 M group (RIP 2017 alignment) along with the CRF02_AG recombinant strain (Blast result). The illustration was created according to HXB2 numbering using the Recombinant Drawing tool available on the Los Alamos HIV sequence Database website *(https://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html)*. The middle panel demonstrate the similarity plot diagram, and the bottom panel shows the bootscan analysis, created by Simplot program vesion 3.5.1 (Lole, Bollinger et al. 1999). The y-axis in the similarity plot indicates the percent identity of the query sequence to a set of reference sequences and the dotted line shows the 90% (significant) similarity; in the bootscan diagram, the y-axis indicates the bootstrap value, and the dotted line indicates 70% (significant) bootstrap value. The x-axis shows the nucleotide position of the HXB2 genome. (B) Indicate a confirmatory analysis by using the same approach as the part (A) of the figure with the only difference being the addition of CRF91_cpx samples to the reference dataset. As indicated by the upper genome diagram, the CRF91_cpx predominates over the rest reference sequences.

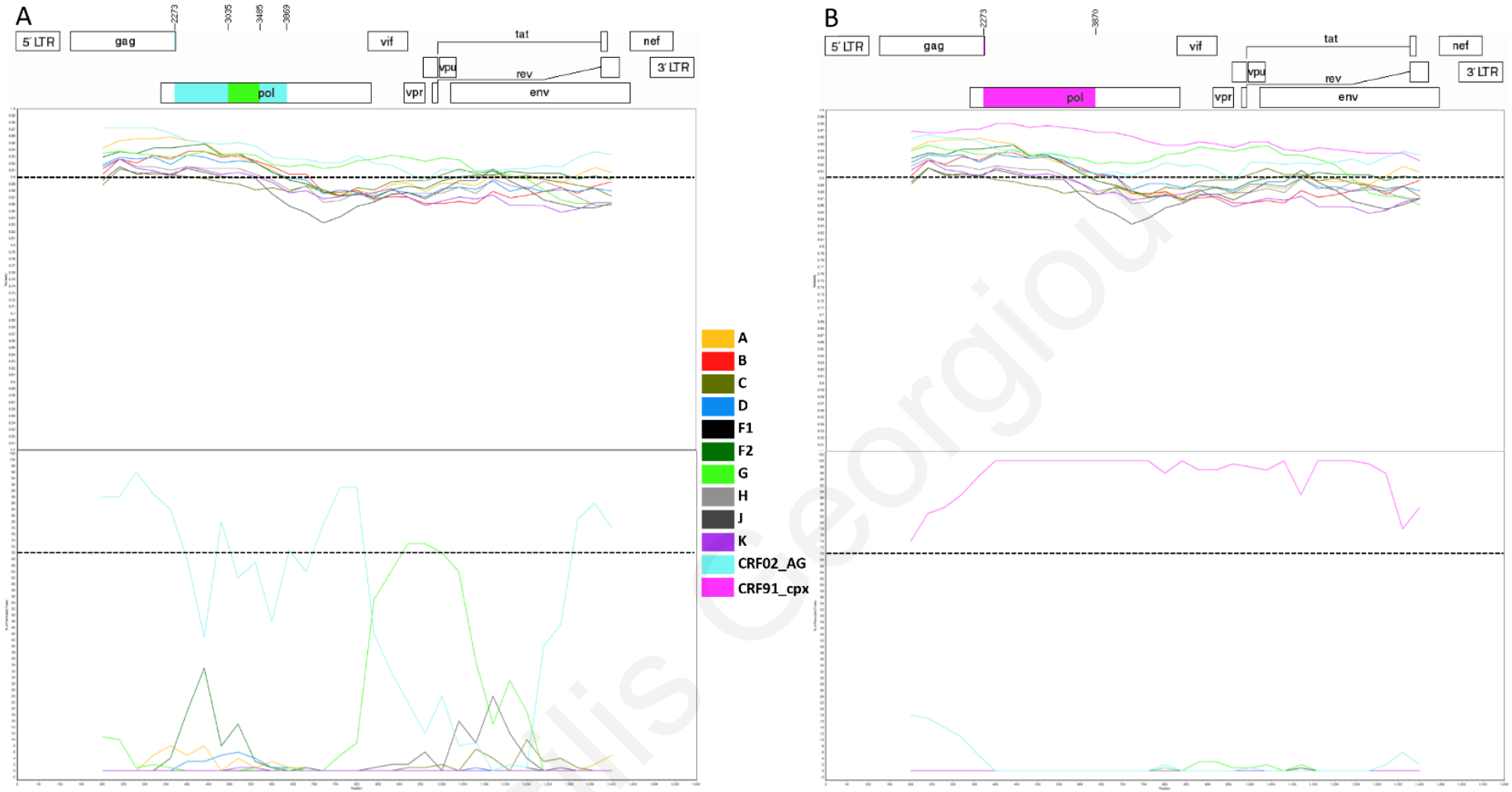


Figure 27: *Identification of CRF91_cpx pol partial genome Circulating in Nigeria (GenBank accession number: MH654876)*. (A) The upper diagram shows the identical recombination breakpoints of the sample MH654876 isolate with the CRF91_cpx mosaic pattern as defined by bioinformatic analysis using as reference dataset the pure subtypes of the HIV-1 M group (RIP 2017 alignment) along with the CRF02_AG recombinant strain (Blast result). The illustration was created according to HXB2 numbering using the Recombinant Drawing tool available on the Los Alamos HIV sequence Database website *(https://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html)*. The middle panel demonstrate the similarity plot diagram, and the bottom panel shows the bootscan analysis, created by Simplot program vesion 3.5.1 (Lole, Bollinger et al. 1999). The y-axis in the similarity plot indicates the percent identity of the query sequence to a set of reference sequences and the dotted line shows the 90% (significant) similarity; in the bootscan diagram, the y-axis indicates the bootstrap value, and the dotted line indicates 70% (significant) bootstrap value. The x-axis shows the nucleotide position of the HXB2 genome. (B) Indicate a confirmatory analysis by using the same approach as the part (A) of the figure with the only difference being the addition of CRF91_cpx samples to the reference dataset. As indicated by the upper genome diagram, the CRF91_cpx predominates over the rest reference sequences.

3.8.2 Recombination Analysis of the Cluster4

Following the same approach used for the recombination analysis of cluster 9 (CRF91_cpx), a maximum likelihood tree was created to identify the uniqueness of the cluster 4 recombinant samples. This analysis was performed using as reference dataset of the RIP alignment (2020) obtained through the Los Alamos HIV sequence Database (www.hiv.lanl.gov/) (Compendium 2018). The RIP 2020 reference dataset contained 198 near full-length genome sequences of all pure group M subtypes including the recently described subtype L along with 118 CRFs, and the N, O and P outgroups. Together with the RIP 2020 alignment, three CRF56_cpx sequences obtained through blast analysis were included. As indicated on the Figure 28, the samples of interest clustered together with branch value of 100/100 (SH-aLRT/ UFBoot). Genetically the most similar reference sequences with the samples of interest were a cluster of CRF56_cpx sequences, where the common ancestral branch of the query sequences with the CRF56_cpx reference sequences was supported by 100/100 (SH-aLRT/ UFBoot) branch value. Subsequently, for further validation of the uniqueness of the recombinant samples of the cluster 4, blast analysis was performed through the NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and the Los Alamos HIV sequence database (https://www.hiv.lanl.gov/content/sequence/BASIC_BLAST/basic_blast.html) in order to find possible identical sequences. The results of the blast analysis reconfirmed the outcome of the phylogenetic analysis revealing a high similarity score with CRF56_cpx sequences. The overall outcome of the aforementioned analysis was that the recombinant samples of cluster 4 present high genetic similarity to the CRF56_cpx strain. Nevertheless, subsequent recombination analysis was performed with the purpose to evaluate the genetic similarity of the recombinant samples of cluster 4 with the CRF56_cpx strain.

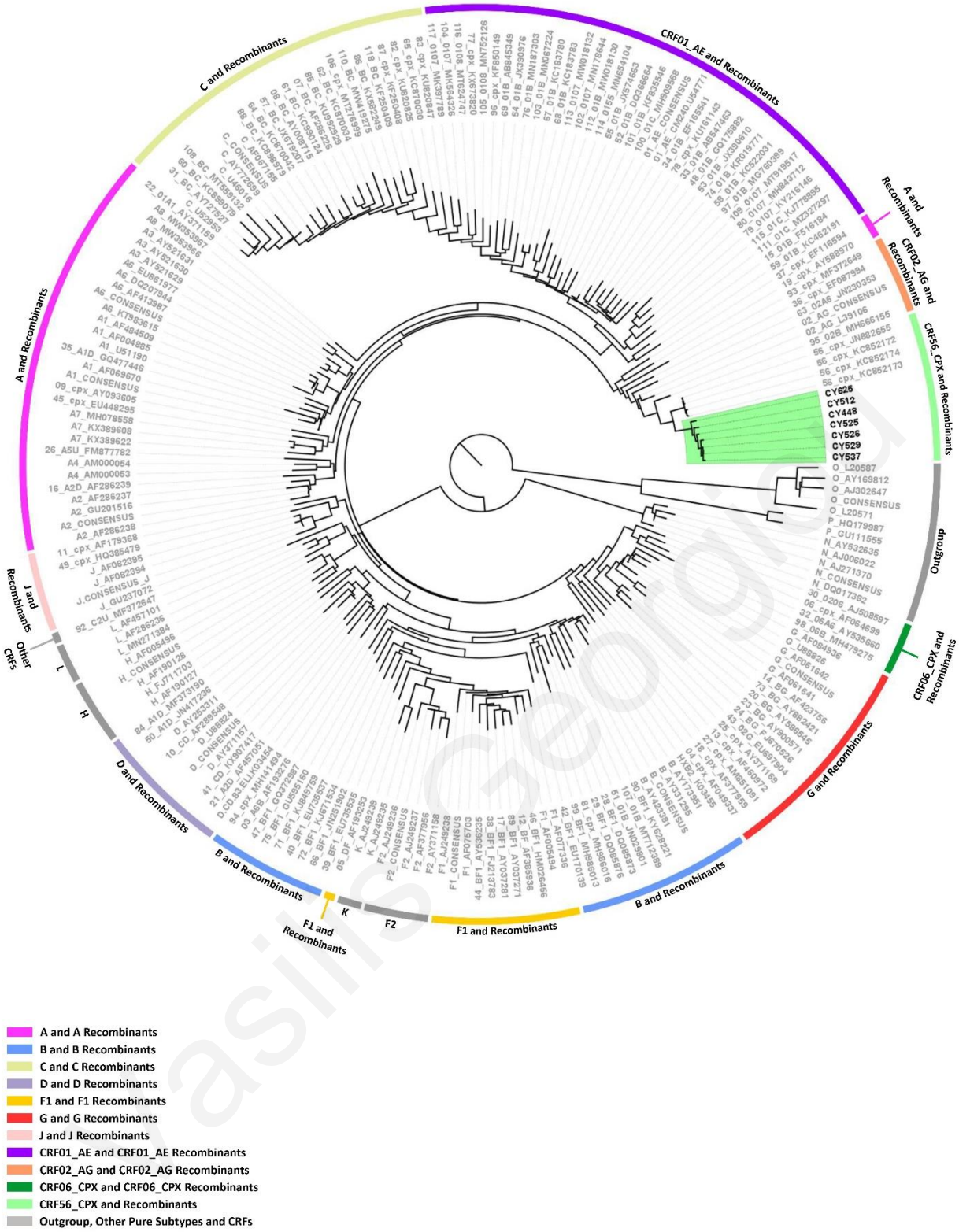


Figure 28: *Maximum likelihood tree analysis of the NFLG sequences of cluster 4 with standard references including the M group pure subtypes and CRFs.* The branch indicated with the green color represents the samples of cluster 4. The outer circular colors indicate the different subtypes and recombinant clusters of the reference dataset used for the creation of the ML tree (RIP 2020 aligment). The cluster that has the closest genetic similarity with the cluster 4 is indicated with the green color and is compose from CRF56_cpx. The gray color represents several clusters of pure subtypes that doesn't predominate in Cyprus, the outgroups and complicated CRFs.

In order to fully examine the genetic similarity of the cluster 4 recombinant samples with the CRF56_cpx recombinant strain as well as to identify the breakpoints and the subtypes that characterize the samples of the cluster 4, all the approaches of the annex 2.11 were applied. However as mention before, only the results of the Simplot program vesion 3.5.1 (Lole, Bollinger et al. 1999) and neighbor joining phylogenetic trees of the representative sample were illustrated as part of this research. The results of the rest of the samples of cluster 4 can be found in the supplementary data (Figure S10-S15).

The results obtained from all approaches indicate that the samples of cluster 4 were characterized by an identical mosaic genomic structure that was composed of the CRF56_cpx, and the subtype G. Specifically the consensus mosaic genomic pattern was composed from three fragments, beginning from the 5' end nucleotide position 790 to 1645 that was characterized by CRF56_cpx with average branch support value of 100% ; from 1646 to 2061, was characterized by subtype G with average branch support value of 75.5%; and from 2062 to 8795, was characterized by CRF56_cpx with average branch support value of 100%.

The Blast analysis that was conducted for each fragment of the genome, reconfirmed that the second fragment (from 1646 to 2061) was characterized with high similarity score with subtype G samples and no similarity with the CRF56_cpx recombinant strain. The overall outcome of the cluster 4 recombination analysis was that there was a high genetic similarity with the CRF56_cpx recombinant strain, however a small part of the genome (from 1646 to 2061) diverges and as such it could be a distinct CRF of CRF56_cpx, G.

Following the recombination analysis the sequences along with the mosaic genome and breakpoints of the samples will be sent to the Los Alamos HIV sequence Database (www.hiv.lanl.gov/) (Compendium 2018). After an independent analysis that will be conducted by the Los Alamos HIV sequence Database experts, if their results came with an agreement with the aforementioned results then the novel CRF of CRF56_cpx, G will be named according to the nomenclature of HIV-1 (Robertson, Anderson et al. 2000).

Sample CY448

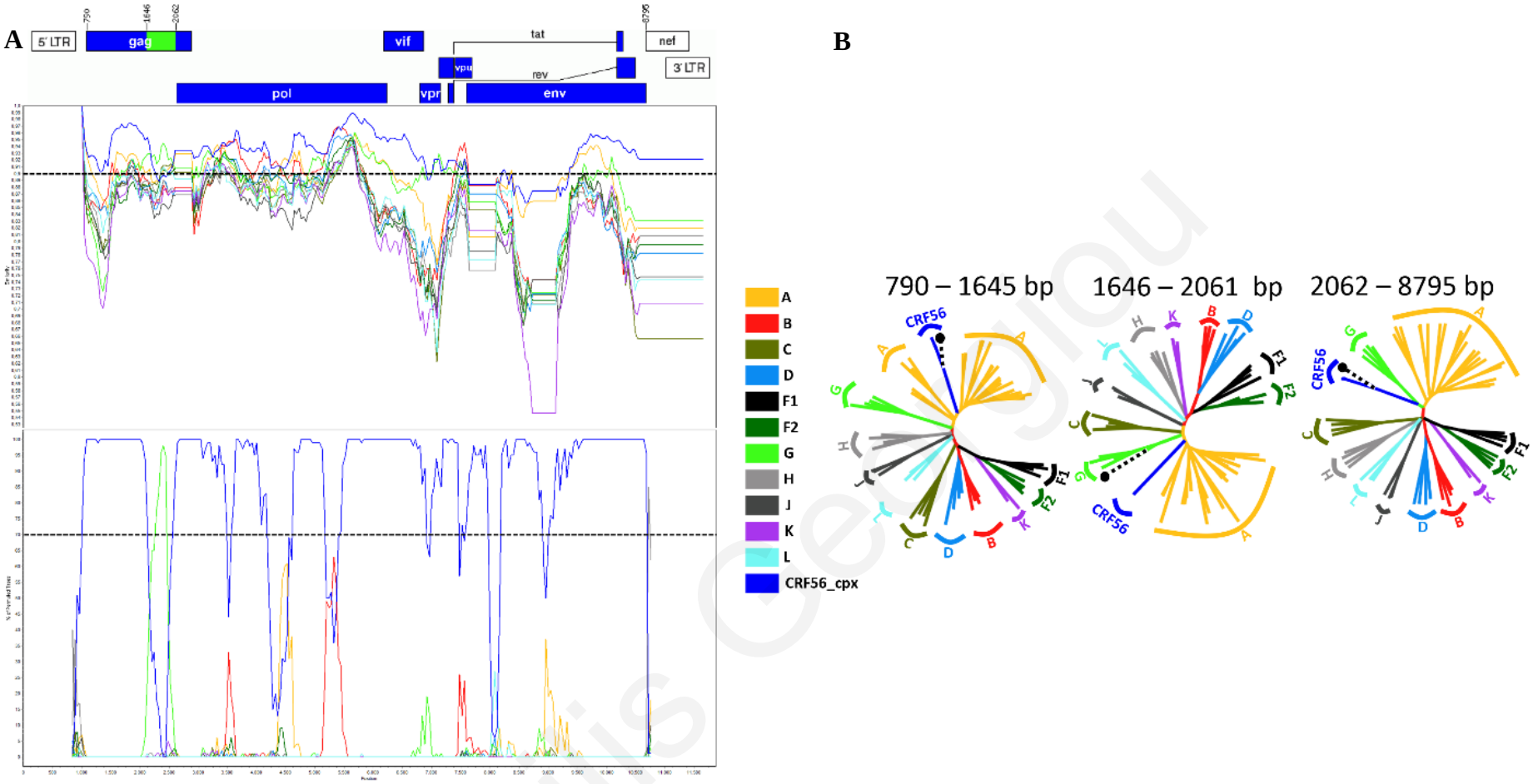


Figure 29: *Recombination analysis of the cluster 4 representative sample CY448.* The left part of the figure (A) represents the mosaic structure of the near full-length genome of the CY448 isolate in comparison to reference strains of all HIV-1 pure subtypes of the M group along with the CRF56_cpx recombinant strain. The upper diagram shows the gene regions and the recombination breakpoints of the CY448 isolate as defined by bioinformatic analysis. The illustration was created according to HXB2 numbering using the Recombinant Drawing tool available on the Los Alamos HIV sequence Database website *(https://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html)*. The middle panel demonstrate the similarity plot diagram, and the bottom panel shows the bootscan analysis, created by Simplot program vesion 3.5.1 (Lole, Bollinger et al. 1999). The y-axis in the similarity plot indicates the percent identity of the query sequence to a set of reference sequences and the dotted line shows the 90% (significant) similarity; in the bootscan diagram, the y-axis indicates the bootstrap value, and the dotted line indicates 70% (significant) bootstrap value. The x-axis shows the nucleotide position of the HXB2 genome. The right part of the Figure (B) indicates the phylogenetic analysis of the inter-breakpoint segments comprises the CY448 strain as defined by the similarity plot and the bootscan analysis. Analysis was performed using the NJ method with the Kimura's two-parameter parameter distance estimation method and bootstrap analysis of 1000 replicates. All reference HIV-1 pure subtypes and recombinants of the M group that were used to construct the trees are denoted with different colors. All the inter-breakpoint segments results are supported by bootstrap value ≥70%; Above of each NJ tree there is the loci of genome segments based on the HXB2 numbering. The reference sequences of the pure subtype were obtained by the RIP 2020 alignment and the CRF56_cpx reference samples were obtained through blast analysis.

3.8.3 Recombination Analysis of Cluster 5 and Cluster 16

The aforementioned recombination analysis was conducted for the clusters 5 and 16 as well. Based on the monthly phylogenetic analysis of the pol region (Figure 19), it was already known that cluster 5 and cluster 16 were genetically similar, however they were characterized by genetic differences sufficient enough to divide them into two distinct clusters. Thus, a maximum likelihood tree was generated to identify the uniqueness of the cluster 5 and 16 recombinant samples. This analysis was performed using as a reference dataset the RIP 2020 alignment obtained through the Los Alamos HIV sequence Database (www.hiv.lanl.gov/) (Compendium 2018). Together with the RIP 2020 alignment, the sample USA_5112 that was obtained through blast analysis was also included into the reference dataset. As indicated on the Figure 30 and Figure 31 the samples of interest cluster together with branch value of 100/100 (SH-aLRT/ UFBoot). Genetically the most similar reference sequences with the samples of interest were a cluster of subtype A and A recombinants sequences. Subsequently, for further validation of the uniqueness of the recombinant samples of the cluster 5 and 16, blast analyses were performed through the NCBI (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and the Los Alamos HIV sequence database (https://www.hiv.lanl.gov/content/sequence/BASIC_BLAST/basic_blast.html) in order to find possible identical sequences. The results of the blast analysis indicate the existence of one uncharacterized sequence (USA_5112, GenBank accession number: MW063005) that had 96% similarity score with the samples of cluster 5 and 94% similarity score with the samples of cluster 16. The overall outcome of the aforementioned analysis was that the recombinant samples of cluster 5 and 16 didn't present high genetic similarity with any of the known HIV-1 pure subtypes or CRFs.

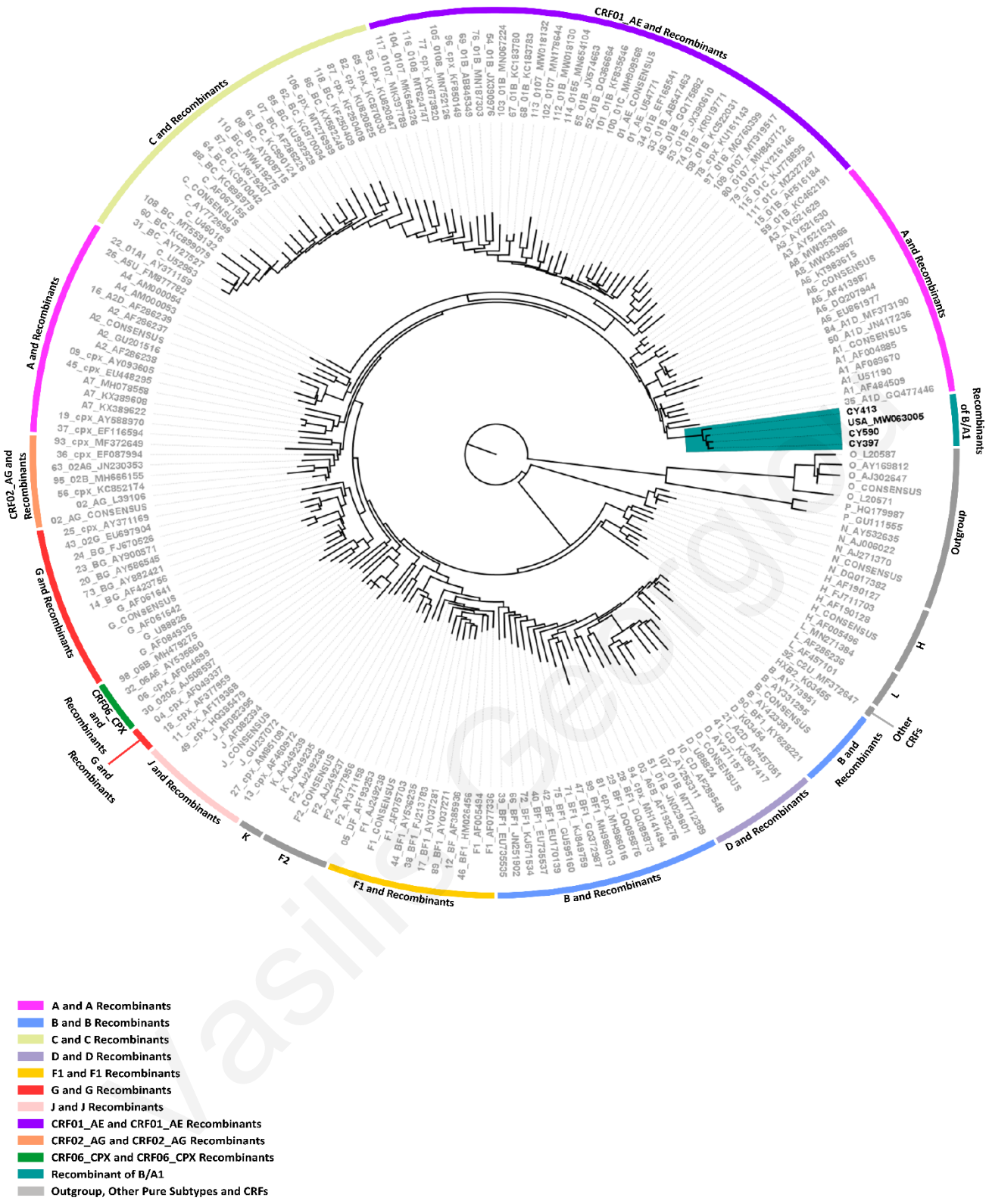


Figure 30: *Maximum* l*ikelihood tree analysis of the NFLG sequences of cluster 5 with standard references including the M group pure subtypes and CRFs.* The branch indicated with the blue/green color represents the samples of cluster 5. The outer circular colors indicate the different subtypes and recombinant clusters of the reference dataset used for the creation of the ML tree (RIP 2020 aligment). The cluster that has the closest genetic similarity with the cluster 5 is indicated with the pink color and is compose from subtype A and A recombinant sequences. The gray color represents several clusters of pure subtypes that doesn't predominate in Cyprus, the outgroups and complicated CRFs.

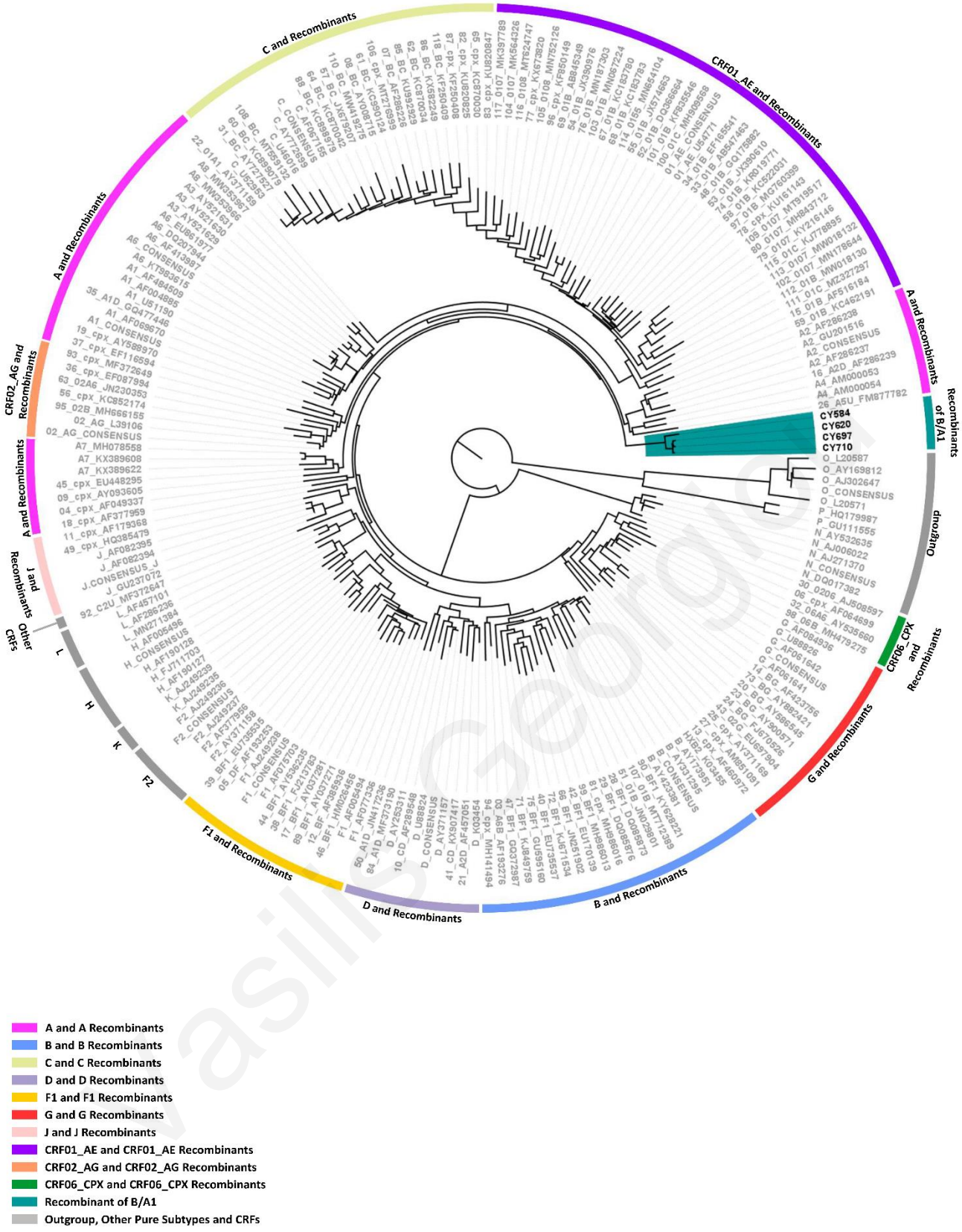


Figure 31:*Maximum likelihood tree analysis of the NFLG sequences of cluster 16 with standard references including the M group pure subtypes and CRFs.* The branch indicated with the blue/green color represents the samples of cluster 16. The outer circular colors indicate the different subtypes and recombinant clusters of the reference dataset used for the creation of the ML tree (RIP 2020 aligment). The cluster that has the closest genetic similarity with the cluster 16 is indicated with the pink color and is compose from subtype A and A recombinant sequences. The gray color represents several clusters of pure subtypes that doesn't predominate in Cyprus, the outgroups and complicated CRFs.

In order to fully examined the genomic mosaic pattern of the recombinant clusters 5 and 16, to evaluate the genetic similarity between the cluster 5 and 16 and to identify if the sample USA_5112 had the same genomic mosaic structure with either the cluster 5 or cluster 16, all approaches of the annex 2.11 were applied. In order to avoid repetitive information only a representative sample for each of the cluster 5 and 16, along with the sample USA_5112 were illustrated as part of this research. The results of the rest of the samples of cluster 5 and 16 can be found in the supplementary data (Figure S16 - S20).

The results obtained from all approaches agree with the phylogenetic analysis that was previously performed, indicating that the samples of cluster 5 and 16 have a similar mosaic pattern and are characterized by the same subtypes, subtype A1 and B, however they presented minor differences in terms of their breakpoints.

Specifically the consensus mosaic genomic pattern of cluster 5 was composed of eight fragments, beginning from the 5' end nucleotide position 790 to 2284 that was characterized by subtype A1 with average branch support value of 100%; from 2285 to 3534, was characterized by subtype B with average branch support value of 96%; from 3535 to 4259, was characterized by subtype A1 with average branch support value of 100%; from 4260 to 4898, was characterized by subtype B with average branch support value of 90.3%; from 4899 to 6054, was characterized by subtype A1 with average branch support value of 98%; from 6055 to 6084, was the recombination range between the subtypes A1 and B; from 6085 to 6321, was characterized by subtype B with average branch support value of 83%; from 6322 to 8485, was characterized by subtype A1 with average branch support value of 100%; and from 8486 to 8690, was characterized by subtype B with average branch support value of 86% (Figure 32). The recombination analysis of the sample USA_5112 revealed an identical recombination pattern as the samples of the cluster 5 (Figure 33).

The Blast analysis that was conducted for each fragment of the genome reconfirmed the aforementioned results indicating that each fragment is either A1 or B subtype.

Next the recombination analysis of the samples of cluster 16 was performed. Specifically the consensus mosaic genomic pattern of cluster 16 was compose from eight fragments, beginning from the 5' end nucleotide position 790 to 2284 that was characterized by subtype A1 with average branch support value of 100%; from 2285 to 3826, was characterized by subtype B with average branch support value of 99%; from 3827 to 4250, was characterized by subtype A1 with average

branch support value of 100%; from 4251 to 4822, was characterized by subtype B with average branch support value of 90.25%; from 4823 to 6084, was characterized by subtype A1 with average branch support value of 97.75%; from 6085 to 6523, was characterized by subtype B with average branch support value of 96%; from 6524 to 8485, was characterized by subtype A1 with average branch support value of 99.75%; and from 8486 to 8690, was characterized by subtype B with average branch support value of 88.5% (Figure 34).

The Blast analysis that was conducted for each fragment of the genome separately reconfirmed the aforementioned results, indicating that each fragment is either A1 or B subtype.

Finally in order to reveal if the minor breakpoint differences between the cluster 5 and cluster 16 mosaic pattern was sufficient to discriminate them as different CRFs, a bootscan and similarity plot analysis was performed using the USA sample 5112 as a query sequence. As indicated in the Figure 35 those minor breakpoints differences are sufficient to indicate that the USA sample 5112 is genetically identical with the cluster 5 and not with the cluster 16.

Following the recombination analysis, the sequences along with the mosaic genomic maps and breakpoints of the samples will be sent to the Los Alamos HIV sequence Database (www.hiv.lanl.gov/) (Compendium 2018). After an independent analysis that will be conducted by the Los Alamos HIV sequence Database experts, if their results came with an agreement with the aforementioned results then the two novel CRFs of A1, B will be named according to the nomenclature of HIV-1 (Robertson, Anderson et al. 2000).

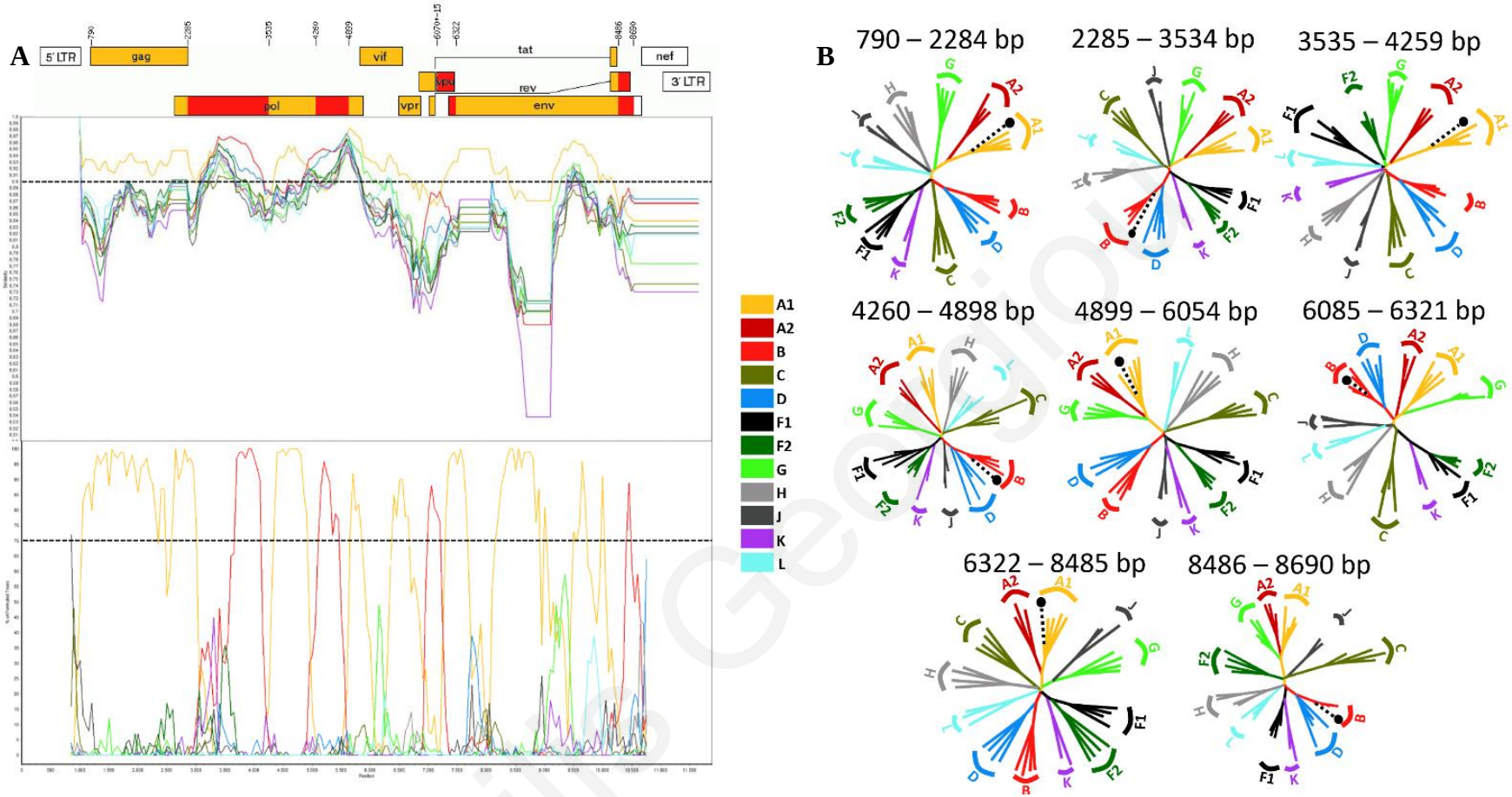


Figure 32: *Recombination analysis of the cluster 5 representative sample CY397.* The left part of the figure (A) represents the mosaic structure of the near full-length genome of the CY397 isolate in comparison to reference strains of all HIV-1 pure subtypes of the M group. The upper diagram shows the gene regions and the recombination breakpoints of the CY397 isolate as defined by bioinformatic analysis. The illustration was created according to HXB2 numbering using the Recombinant Drawing tool available on the Los Alamos HIV sequence Database website *(https://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html)*. The middle panel demonstrate the similarity plot diagram, and the bottom panel shows the bootscan analysis, created by Simplot program vesion 3.5.1 (Lole, Bollinger et al. 1999). The y-axis in the similarity plot indicates the percent identity of the query sequence to a set of reference sequences and the dotted line shows the 90% (significant) similarity; in the bootscan diagram, the y-axis indicates the bootstrap value, and the dotted line indicates 70% (significant) bootstrap value. The x-axis shows the nucleotide position of the HXB2 genome. The right part of the Figure (B) indicates the phylogenetic analysis of the inter-breakpoint segments comprises the CY397 strain as defined by the similarity plot and the bootscan analysis. The analysis was performed using the NJ method with the Kimura's two-parameter parameter distance estimation method and bootstrap analysis of 1000 replicates. All reference HIV-1 pure subtypes of the M group that were used to construct the trees are denoted with different colors. All the inter-breakpoint segments results are supported by bootstrap value ≥70%; Above of each NJ tree there is the loci of genome segments based on the HXB2 numbering. The reference dataset was consisted of pure subtype obtained by the RIP 2020 alignment along with B subtype sequences obtain through blast.

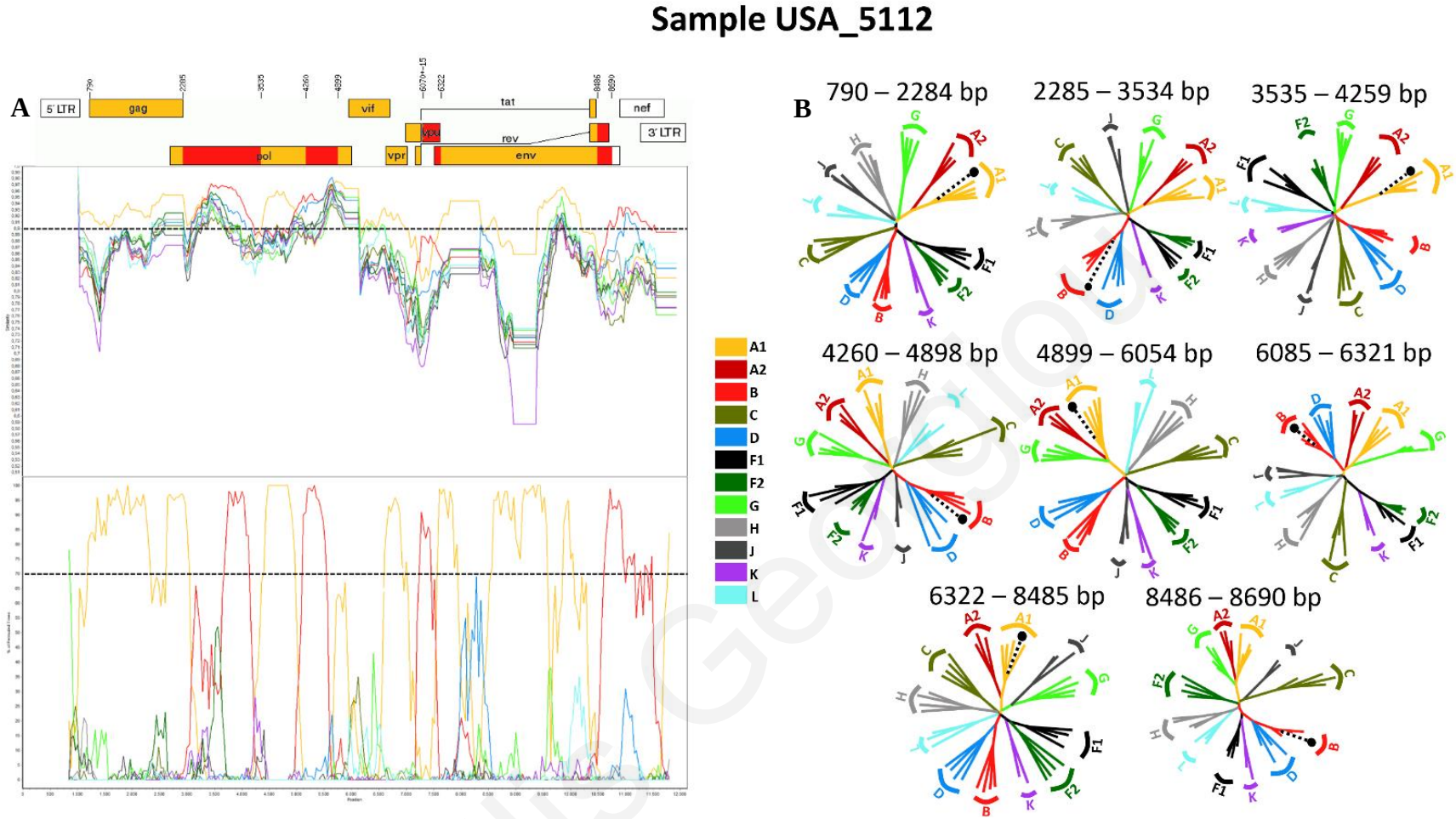


Figure 33: *Recombination analysis of the sample USA_5112.* The left part of the figure (A) represents the mosaic structure of the near full-length genome of the USA_5112 isolate in comparison to reference strains of all HIV-1 pure subtypes of the M group. The upper diagram shows the gene regions and the recombination breakpoints of the USA_5112 isolate as defined by bioinformatic analysis. The illustration was created according to HXB2 numbering using the Recombinant Drawing tool available on the Los Alamos HIV sequence Database website *(https://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html)*. The middle panel demonstrate the similarity plot diagram, and the bottom panel shows the bootscan analysis, created by Simplot program vesion 3.5.1 (Lole, Bollinger et al. 1999). The y-axis in the similarity plot indicates the percent identity of the query sequence to a set of reference sequences and the dotted line shows the 90% (significant) similarity; in the bootscan diagram, the y-axis indicates the bootstrap value, and the dotted line indicates 70% (significant) bootstrap value. The x-axis shows the nucleotide position of the HXB2 genome. The right part of the Figure (B) indicates the phylogenetic analysis of the inter-breakpoint segments comprises the USA_5112 strain as defined by the similarity plot and the bootscan analysis. The analysis was performed using the NJ method with the Kimura's two-parameter parameter distance estimation method and bootstrap analysis of 1000 replicates. All reference HIV-1 pure subtypes of the M group that were used to construct the trees are denoted with different colors. All the inter-breakpoint segments results are supported by bootstrap value ≥70%; Above of each NJ tree there is the loci of genome segments based on the HXB2 numbering. The reference dataset was consisted of pure subtype obtained by the RIP 2020 alignment along with B subtype sequences obtain through blast.

Sample CY584

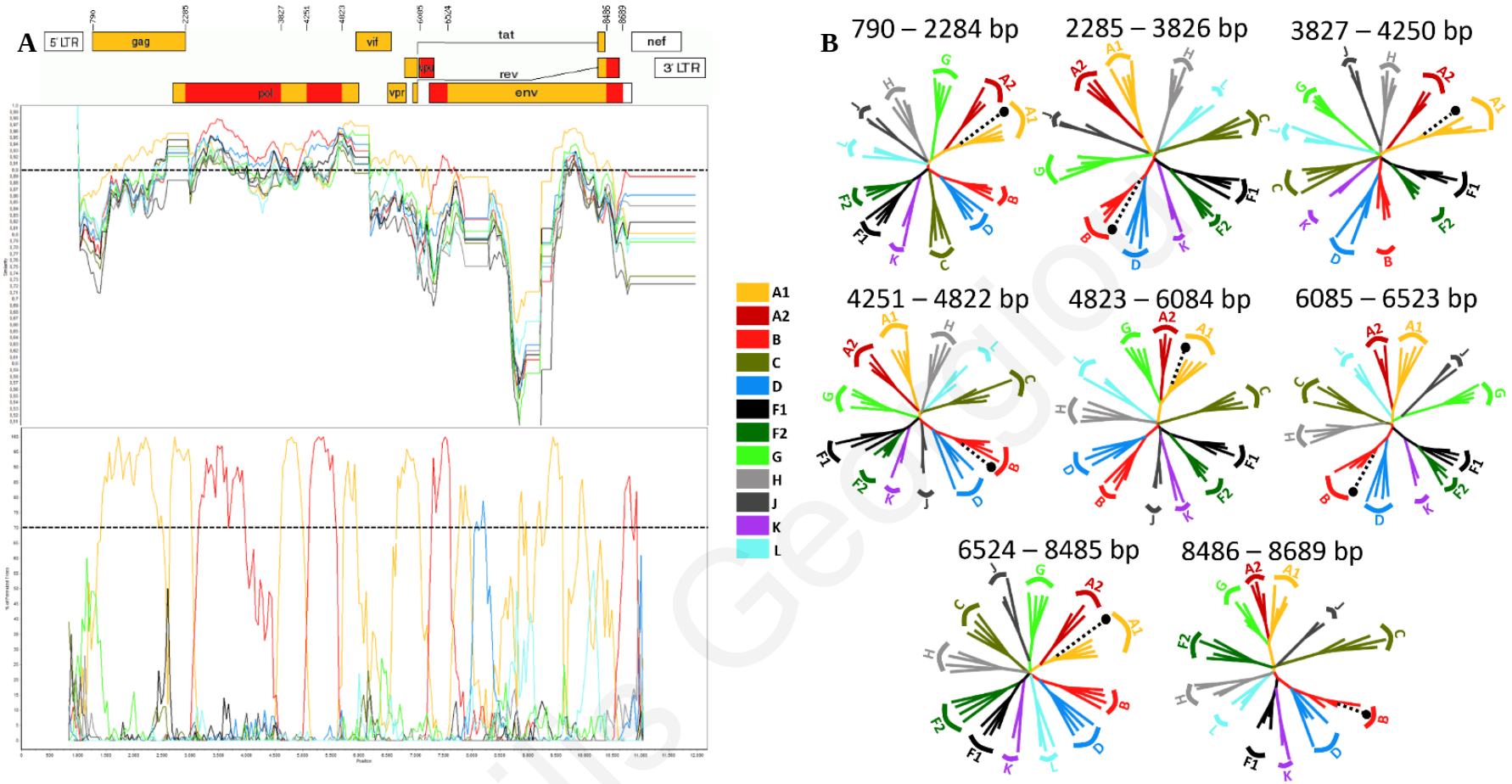


Figure 34: *Recombination analysis of the cluster 16 representative sample CY584.* The left part of the figure (A) represents the mosaic structure of the near full-length genome of the CY584 isolate in comparison to reference strains of all HIV-1 pure subtypes of the M group. The upper diagram shows the gene regions and the recombination breakpoints of the CY584 isolate as defined by bioinformatic analysis. The illustration was created according to HXB2 numbering using the Recombinant Drawing tool available on the Los Alamos HIV sequence Database website *(https://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html)*. The middle panel demonstrate the similarity plot diagram, and the bottom panel shows the bootscan analysis, created by Simplot program vesion 3.5.1 (Lole, Bollinger et al. 1999). The y-axis in the similarity plot indicates the percent identity of the query sequence to a set of reference sequences and the dotted line shows the 90% (significant) similarity; in the bootscan diagram, the y-axis indicates the bootstrap value, and the dotted line indicates 70% (significant) bootstrap value. The x-axis shows the nucleotide position of the HXB2 genome. The right part of the Figure (B) indicates the phylogenetic analysis of the inter-breakpoint segments comprises the CY584 strain as defined by the similarity plot and the bootscan analysis. The analysis was performed using the NJ method with the Kimura's two-parameter parameter distance estimation method and bootstrap analysis of 1000 replicates. All reference HIV-1 pure subtypes of the M group that were used to construct the trees are denoted with different colors. All the inter-breakpoint segments results are supported by bootstrap value ≥70%; Above of each NJ tree there is the loci of genome segments based on the HXB2 numbering. The reference dataset was consisted of pure subtype obtained by the RIP 2020 alignment along with B subtype sequences obtain through blast.

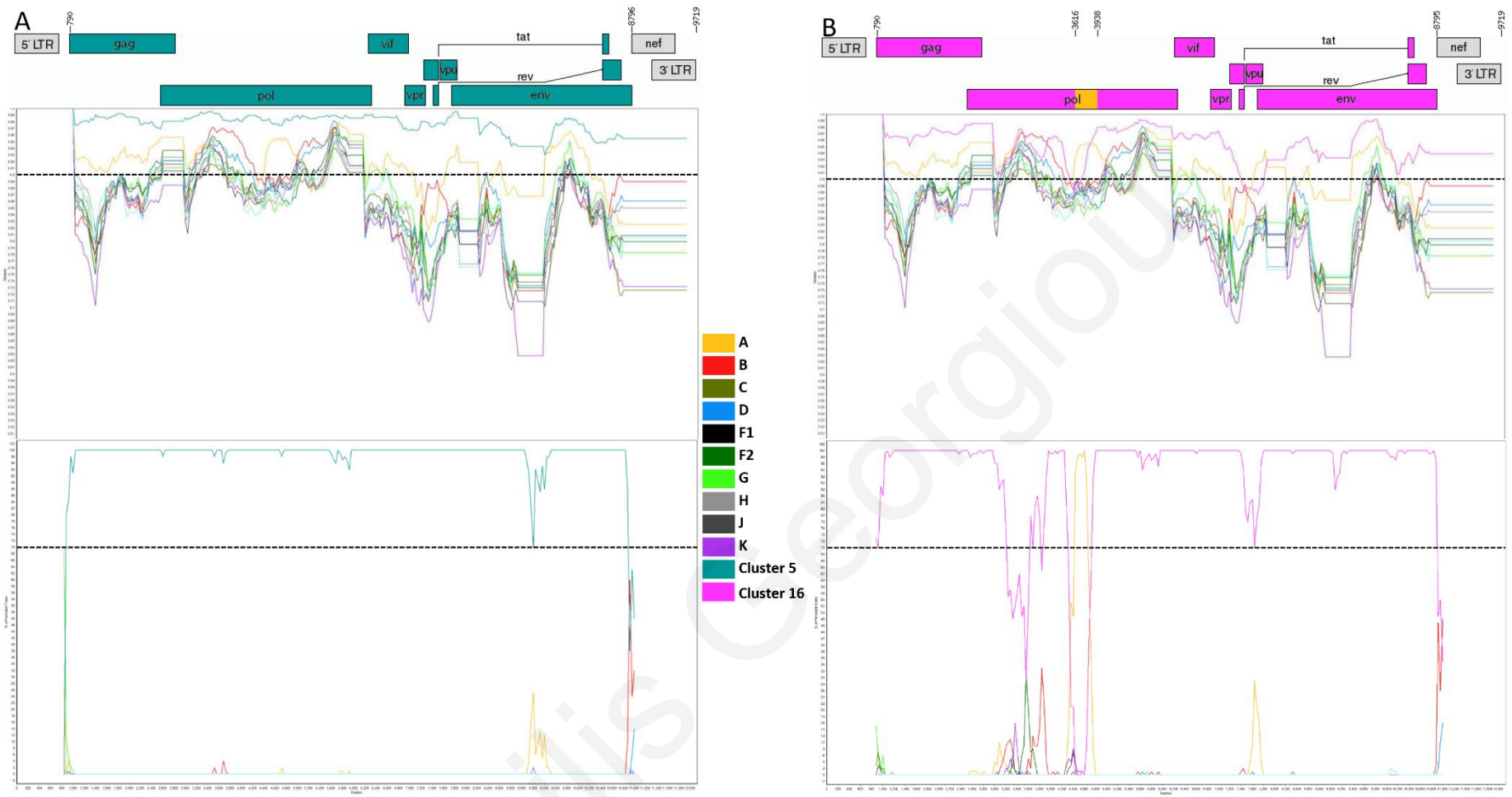


Figure 35: *Comparison of the sample USA_5112 (GenBank accession number: MW063005) with the Cluster 5 and Cluster 16 recombinant samples*. (A) The upper diagram shows the identical genome of the USA_5112 sample isolate with the Cluster 5 using as reference dataset the pure subtypes of the HIV-1 M group (RIP 2020 alignment) along with the samples of cluster 5. The illustration was created according to HXB2 numbering using the Recombinant Drawing tool available on the Los Alamos HIV sequence Database website *(https://www.hiv.lanl.gov/content/sequence/DRAW_CRF/recom_mapper.html)*. The middle panel demonstrate the similarity plot diagram, and the bottom panel shows the bootscan analysis, created by Simplot program vesion 3.5.1 (Lole, Bollinger et al. 1999). The y-axis in the similarity plot indicates the percent identity of the query sequence to a set of reference sequences and the dotted line shows the 90% (significant) similarity; in the bootscan diagram, the y-axis indicates the bootstrap value, and the dotted line indicates 70% (significant) bootstrap value. The x-axis shows the nucleotide position of the HXB2 genome. (B) Comparison of the USA_5112 sample with the cluster 16, using the same approach as the part (A) of the figure. The sample USA_5112 presents a difference at the nucleotide position 3616-3938 compared to the cluster 16 samples indicating that the minor breakpoint differences of the cluster 5 and cluster 16 are sufficient to discriminate a foreign sample.

4.0 DISCUSSION

In general, epidemiological research is focuses on the study of the distribution, the dynamics, and the determinants of diseases and viruses in populations. The risk of viral infection and/or disease in a human population can be characterized as a multivariant equation since it is affected by the characteristics of the virus, the individuals, and the *env*ironmental/ecological factors. Virus epidemiology aims to uncover these factors using numerus experimental phylogenetic and statistical approaches to provide insights for explaining the existence of viral diseases and for managing disease-control measures by using prevention strategies (Burrell, Howard et al. 2017). In terms of the HIV-1, the success of the virus is based on several factors. One crucial factor is the hypervariability of the virus, and as such molecular epidemiology is used throughout the years as a sensitive tool for monitoring the circulation of numerous genetic variants of HIV-1, offering to study the origin and spread of different variants of the virus, in geographical areas, between groups with different risk levels.

Cyprus has been characterized as an extremely interesting model for studying HIV-1 epidemic and transmission, because a complete and densely sampled dataset containing socio-demographic, clinical, virological and behavioral data has been continuously maintained over the years (Pineda-Peña, Theys et al. 2018). Thus, numerus molecular epidemiology studies have been published indicating the polyphyletic HIV-1 infection in Cyprus from 1986 to 2012.

Initially, a study was conducted that examined HIV-1 strains isolated from 77 HIV-1 infected individuals representing 38% of the known infected population in Cyprus in the period 1986 to 2006. Phylogenetic analyses of the obtained viral sequences of this study indicated that subtype B was the predominant subtype (61%), followed by subtype A (23.3%), subtype C (5.2%), CRF02_AG (3.9%), and subtype D, CRF01_AE, and CRF04_cpx (1.3% each) (Kousiappa, Van De Vijver, David AMC et al. 2009). Moving on, another study conducted in Cyprus that examined HIV-1 strains isolated from 74 subjects from 2007 to 2009 revealed that the subtype B remained as the main subtype (48.6%), followed by subtype A (18.9%), subtype C (10.8%), CRF02_AG (8.1%), CRF11_cpx (2.7%), (sub)subtype F1 and CRF37_cpx (1.4% each) (Kousiappa, Achilleos et al. 2011). Next, a study that examined the pol region sequences of HIV-1 strains isolated in Cyprus from drug naïve patients of the 2010-2012 cohort, illustrated that subtypes B and A1 once

again were the most common subtypes, which accounted for 41.0 and 19.0% of the whole cohort, respectively. These were followed by subtype C (7.0%), F1 (8.0%), CRF02_AG (4.0%), A2 (2.0%), other circulating recombinant forms (CRFs) (7.0%) and unique recombinant forms (URFs) (12%) (Kostrikis, Hezka et al. 2018). Reaching today, samples and demographic data collected from 93 consenting newly diagnosed HIV-1 infected patients in Cyprus from October 2019 to December 2020 indicated the existence of 22 different group M subtypes, CRFs and recombinants with the majority to be subtype A1 (n=26, 28%), followed by CRF02_AG (n=15, 16%) and B (n=14, 15%). Moreover, 23% of the samples were recombinants (n=21), while 48% of those were CRF02_AG recombinants (n=10) (19th European Meeting on HIV & Hepatitis HIV Poster 2021 VE Poster 32, Cicek Topcu, data not published).

The numerus studies that have been conducted in Cyprus by the Laboratory of Biotechnology and Molecular Virology of the University of Cyprus, using data from the beginning of the pandemic until today, without any doubt, proved the existence and active growth of a highly polyphyletic HIV-1 epidemic in Cyprus. The existence of numerus subtypes in the same geographical areas, in time, results in simultaneous infection of individuals with viruses belonging to different subtypes (Yaseen, Abuharfeil et al. 2017). As a result, due to recombination events (detail description in annex 1.7.1), novel HIV-1 recombinant strains arise over time.

In this study, the near full-length genome (NFLG) sequences from 24 isolates were successfully amplified and sequenced. The NFLG sequences of those samples were spanning from gag gene to the env gene corresponding to the location 790-8795 bp of the HXB2 reference strain (GenBank entry number: K03455). Phylogenetic analyses revealed the existence of four different recombinant clusters (cluster 9 “Rec. of 02_AG, G, J”;cluster 4“Rec of CRF56_cpx, G”; and clusters 5 and 16 “Rec of A1, B”) that did not cluster with any known HIV-1 subtype or CRF, and formed a distinct monophyletic branch with a bootstrap value of 100% (Figure 22, Figure 28, Figure 30 and Figure 31), indicating potential novel CRFs in Cyprus. Furthermore, similarity and bootscan analyses revealed the same mosaic pattern for the sequences in each of the corresponding clusters, revealing six putative inter-subtype recombination breakpoints for the cluster 9 “Rec. of 02_AG, G, J”, two inter-subtype recombination breakpoints for the cluster 4 “Rec of CRF56_cpx, G”, and seven inter-subtype recombination breakpoints for the two recombinant clusters 5 and 16 “Rec of A1, B”. Similar results were also obtained using RIP and jpHMM analyses, NCBI

genotyping tool and REGA HIV-1 Subtyping Tool. Subregion neighbor-joining tree analyses further confirmed the breakpoints of the four recombinant clusters.

Specifically, for the cluster 9 "Rec. of 02_AG,G,J" the following breakpoints and genome fragments were identified: I (790–3034 nt) CRF02_AG, II (3035–3082 nt) recombination range between the subtypes CRF02_AG and subtype G-like, III (3083–3410 nt) CRF02_AG, IV (3411–3484 nt) recombination range between the subtypes G and CRF02_AG, V (3485-5421 nt) CRF02_AG, VI (5422-6622 nt) subtype G, VII (6623-7553 nt) undefined region, VIII (7554-8433) subtype G and IX (8434-8686) subtype J-like (Figure 24). Taken together, phylogenesis and recombinant structures of these sequences were distinct from any known CRF reported to date, and the strains were isolated from ten HIV-1 infected patients without any epidemiological linkage in Cyprus, meeting the criteria for designation of a new CRF. Therefore, these new recombinants were characterized as a novel circulating recombinant form (CRF91_cpx) by the Los Alamos HIV Database according to the standards of HIV-1 nomenclature (Robertson, Anderson et al. 2000). It is important to note that since the identification of the CRF91_cpx, two more patient samples from the Cyprus cohort have been introduced into the CRF91_cpx transmission cluster, demonstrating an active growth. Additionally blast analysis revealed high similarity score with two partial pol genomes derived from Nigeria (GenBank accession numbers: MH654941and MH654876). Similarity, bootscan and subregion maximum likelihood tree analyses indicated that those samples are characterized by the same subtypes and breakpoints reconfirming that they are partial CRF91_cpx sequences derived from Nigeria (Figure 26 and Figure 27).

Subregion Blast analysis that was performed indicated that the last fragment (from 8434 to 8686 nt) was characterized by high similarity score with CRF06_cxp samples. Thus, a second phylogenetic analysis was performed for the last fragment (8434-8686) revealing a 100% average brunch support value with the CRF06_cpx reference sequences. It is important to note that CRF06_cpx at the nucleotide position 8434 to 8686 is also characterized by the subtype J and as such the last fragment was determine as J-like subtype (Figure 25)

During the investigation of the genome mosaic pattern analysis of the cluster 9 one particular sample (CY467) seemed to diverge from the rest of the samples. This result came with an agreement with the pol region maximum likelihood tree analysis, since CY467 was included into the cluster 9, however it was characterized with higher genetic difference compared to the rest of the samples of the cluster 9 (Figure 19). Further similarity, bootscan and neighbor-joining tree

analyses revealed that CY467 was characterized by a CRF91_cpx backbone with the only difference being one fragment on the genome that was characterized by subtype B (from 3804 to 4280 nt). Thus, this sample was proved to be a URF of CRF91_cpx, B (Figure 23).

Moving on, the same strategy was applied to the rest of the recombinant clusters (cluster 4 "Rec of CRF56_cpx, G"; and clusters 5 and 16 "Rec of A1, B"). In particular for the cluster 4 the initial maximum likelihood tree analysis indicated that those seven samples were genetically similar to the CRF56_cpx strain (Figure 28). The blast analysis that was followed reconfirmed the genetic similarity of the samples of the cluster 4 with the CRF56_cpx reference strains. Nevertheless, for the purpose of an exhaustive examination, similarity, bootscan and neighbor-joining tree analysis were performed. The comprehensive results of all the approaches indicated that the samples of the cluster 4 were characterized by a CRF56_cpx backbone with the only difference being one fragment located in the gag region that was characterized by the subtype G (from 1641 to 2061 nt) (Figure 29). The difference of only 420 nucleotides may seem minor, however throughout the years several CRFs have been identified that share the same characteristic (Zhang, Feng et al. 2019, Feng, Yue, Zhang et al. 2018, Li, X., Ning et al. 2013). Taken together, all the evidence indicated that the cluster 4 samples are distinct from any known CRFs reported to date, and the sequences were isolated from seven HIV-1 infected patients without any epidemiological linkage in Cyprus, meeting the criteria for designation of a new CRF. Therefore, the mosaic genome along with the sequences of the cluster 4 samples will be sent to the Los Alamos HIV-1 sequence database with the aim to claim the novel CRF if the evidence meets the criteria.

Finally, the recombination analyses of the cluster 5 and 16 were performed. As indicated on the maximum likelihood tree of the pol region (Figure 19) these two clusters were genetically very similar, however, they are characterized by some genetic differences that led to the formation of two separate distinct clusters. The analyses performed agreed with the aforementioned results, showing that the two clusters shared the same subtypes with the same order on their mosaic genome, but they are characterized by slightly different recombination breakpoints. In particular, for the cluster 5 "Rec. of B,A1" the following breakpoints and genome fragments were identified: I (790–2284 nt) subtype A1, II (2285–3534 nt) subtype B, III (3535–4259 nt) subtype A1, IV (4260–4898 nt) subtype B, V (4899-6054 nt) subtype A1, VI (6055-6084 nt) recombination range between the subtypes A1 and B, VII (6085-6321 nt) subtype B, VIII (6322-8485) subtype A1 and

IX (8486-8690) subtype B (Figure 32). The cluster 16 mosaic genome is characterized by the same fragments and subtypes with the cluster 5 with the only noticeable differences to be at the nucleotide positions 3535-3827 and 6322-6524. In these regions, the cluster 5 was characterized by subtype A1 and the cluster 16 was characterized by subtype B, forming a total of 494 nucleotides that diverge in their genome (Figure 32, Figure 34 and S21). The blast analysis that was followed reconfirmed the genetic similarity between the clusters 5 and 16 since both clusters indicated high similarity score (cluster 5, 96% and cluster 16, 94%) with the same sample that was derived from United States of America (USA_5112, GenBank accession number: MW063005). In order to identify if the USA_5112 sample belonged to any of those two clusters, a comprehensive recombination analysis was performed, which indicated that this sample shared the exact same mosaic genome with the cluster 5 (Figure 33). Furthermore, taking advantage of the aforementioned results, the USA_5112 sample was used in order to examine if the minor differences between clusters 5 and 16 are sufficient for the categorization of an outgroup sample into one of those two clusters. As indicated in the Figure 35, the genetic differences between the clusters 5 and 16 are indeed sufficient to categorize an outgroup sample. Taken together, all the evidence indicated that the clusters 5 and 16 are distinct between them and from any known CRF reported to date, and the sequences were isolated from six HIV-1 infected patients without any epidemiological linkage in Cyprus, meeting the criteria for designation of a new CRF. Therefore, the mosaic genomic map along with the sequences of the clusters 5 and 16 samples will be sent to the Los Alamos HIV-1 sequence database to claim another two new CRFs, if the evidence meets the criteria.

In summary, we characterized a novel CRF91_cpx strain and a URF of CRF91_cpx with subtype B and we also presented evidence of three additional putative CRFs. The present findings highlight the urgent need for continuous molecular screening and epidemic surveillance to implement effective measures to reduce HIV-1 transmission. The results of this study provide new insights and supply valuable data that aid understanding the influx of new strains that shape the polyphyletic infection of HIV-1 in Cyprus. This study provides the groundwork that facilitate the reconnaissance of new HIV-1 recombinants, consequently aiding in the research of HIV, and safeguarding public health.

With the aim of the comprehensive understanding of the newly identified CRFs in Cyprus, future studies may include the implementation of phylodynamic approaches with the purpose to

reveal the time of origin and the worldwide distribution of the CRF91_cpx along with the three putative CRFs.

Collectively, HIV-1 recombinants are increasingly prominent in global and regional HIV epidemics, which has important implications in various perspectives including transmission routes, disease progression, drug resistance and response to HAART, viral load measurement, diagnostics, pathogenesis, immune response, and vaccine development. As such the continuation and improvement of the surveillance of the global molecular epidemiology of HIV-1 is crucial for safeguarding the global health.

SUPPLEMENTARY MATERIALS

The following are available here, Figure S1: Recombination analysis of the CRF91_cpx, Sample CY520; Figure S2: Recombination analysis of the CRF91_cpx, Sample CY533; Figure S3: Recombination analysis of the CRF91_cpx, Sample CY614; Figure S4: Recombination analysis of the CRF91_cpx, Sample CY622; Figure S5: Recombination analysis of the CRF91_cpx, Sample CY630; Figure S6: Recombination analysis of the CRF91_cpx, Sample CY640; Figure S7: Recombination analysis of the CRF91_cpx, Sample CY670; Figure S8: Recombination analysis of the CRF91_cpx, Sample CY686; Figure S9: Recombination analysis of the CRF91_cpx, Sample CY742; Figure S10: Recombination analysis of the cluster 4, sample CY512; Figure S11: Recombination analysis of the cluster 4, sample CY525; Figure S12: Recombination analysis of the cluster 4, sample CY526; Figure S13: Recombination analysis of the cluster 4, sample CY529; Figure S14: Recombination analysis of the cluster 4, sample CY537; Figure S15: Recombination analysis of the cluster 4, sample CY625; Figure S16: Recombination analysis of the cluster 5, sample CY413; Figure S17: Recombination analysis of the cluster 5, partial sample CY590; Figure S18: Recombination analysis of the cluster 16, sample CY620; Figure S19: Recombination analysis of the cluster 16, sample CY697; Figure S20: Recombination analysis of the cluster 16, sample CY710; Figure S21: Comparison of the consensus mosaic genomes of the clusters 5 and 16; Table S1: RIP 2020 alignment used for all ML Trees; Table S2: NJ Reference Dataset used for CRF91_cpx; Table S3: NJ Reference Dataset used for cluster 4; Table S4: NJ Reference Dataset used for clusters 5 and 16.

FUNDING

This work was co-funded by the European Regional Development Fund and the Republic of Cyprus through the Research and Innovation Foundation and the Structural Funds 2021-2027 (Project: EXCELLENCE/0918/0093), the Ministry of Health of the Republic of Cyprus (Contract No 240/2019) and the University of Cyprus through internal funds awarded to Prof. Leondios G. Kostrikis.

DECLARATION OF COMPETING INTEREST

The author declares no competing financial interests.

SEQUENCES DATA

The gene sequences of the cluster 9 (CRF91_cpx/URF of CRF91,B) were deposited in the GenBank with the following accession numbers: CY467, OK584018; CY494, OK283056; CY520, OK283057; CY533, OK283058; CY614, OK283059; CY622, OK283060; CY630, OK283061; CY640, OK283062; CY670, OK283063; CY686, OK283064; CY742, OK283065.The samples of the three putative CRFs weren't deposited into the GenBank since the approval of the Los Alamos HIV-1 Sequence Database is pending.

BIBLIOGRAPHY

ALAEUS, A., LIDMAN, K., BJÖRKMAN, A., GIESECKE, J. and ALBERT, J., 1999. Similar rate of disease progression among individuals infected with HIV-1 genetic subtypes AD. *Aids,* **13**(8), pp. 901-907.

ALTMAN, L.K., 1982. New homosexual disorder worries health officials. *The New York Times,* **11**, pp. C1-6.

AMORNKUL, P.N., KARITA, E., KAMALI, A., RIDA, W.N., SANDERS, E.J., LAKHI, S., PRICE, M.A., KILEMBE, W., CORMIER, E. and ANZALA, O., 2013. Disease progression by infecting HIV-1 subtype in a seroconverter cohort in sub-Saharan Africa. *AIDS (London, England),* **27**(17), pp. 2775.

AMORNKUL, P.N., TANSUPHASAWADIKUL, S., LIMPAKARNJANARAT, K., LIKANONSAKUL, S., YOUNG, N., EAMPOKALAP, B., KAEWKUNGWAL, J., NAIWATANAKUL, T., VON BARGEN, J. and HU, D.J., 1999. Clinical disease associated with HIV-1 subtype B ,_and E infection among 2104 patients in Thailand. *Aids,* **13**(14), pp. 1963-1969.

APETREI, C., DESCAMPS, D., COLLIN, G., LOUSSERT-AJAKA, I., DAMOND, F., DUCA, M., SIMON, F. and BRUN-VÉZINET, F., 1998. Human immunodeficiency virus type 1 subtype F reverse transcriptase sequence and drug susceptibility. *Journal of virology,* **72**(5), pp. 3534-3538.

ARTS, E.J. and HAZUDA, D.J., 2012. HIV-1 antiretroviral drug therapy. *Cold Spring Harbor perspectives in medicine,* **2**(4), pp. a007161.

AUTRAN, B., CARCELAIN, G., LI, T.S., BLANC, C., MATHEZ, D., TUBIANA, R., KATLAMA, C., DEBRE, P. and LEIBOWITCH, J., 1997. Positive effects of combined antiretroviral therapy on CD4 T cell homeostasis and function in advanced HIV disease. *Science,* **277**(5322), pp. 112-116.

BAETEN, J.M., CHOHAN, B., LAVREYS, L., CHOHAN, V., MCCLELLAND, R.S., CERTAIN, L., MANDALIYA, K., JAOKO, W. and JULIE, O., 2007. HIV-1 subtype D infection is associated with faster disease progression than subtype A in spite of similar plasma HIV-1 loads. *The Journal of infectious diseases,* **195**(8), pp. 1177-1180.

BAROUCH, D.H., 2008. Challenges in the development of an HIV-1 vaccine. *Nature,* **455**(7213), pp. 613-619.

BAROUCH, D.H. and KORBER, B., 2010. HIV-1 vaccine development after STEP. *Annual Review of Medicine,* **61**, pp. 153-167.

BAROUCH, D.H., TOMAKA, F.L., WEGMANN, F., STIEH, D.J., ALTER, G., ROBB, M.L., MICHAEL, N.L., PETER, L., NKOLOLA, J.P. and BORDUCCHI, E.N., 2018. Evaluation of a mosaic HIV-1 vaccine in a multicentre, randomised, double-blind, placebo-controlled, phase 1/2a clinical trial (APPROACH) and in rhesus monkeys (NHP 13-19). *The Lancet*, **392**(10143), pp. 232-243.

BARRÉ-SINOUSSI, F., CHERMANN, J., REY, F., NUGEYRE, M.T., CHAMARET, S., GRUEST, J., DAUGUET, C., AXLER-BLIN, C., VÉZINET-BRUN, F. and ROUZIOUX, C., 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, **220**(4599), pp. 868-871.

BEKKER, L., MOODIE, Z., GRUNENBERG, N., LAHER, F., TOMARAS, G.D., COHEN, K.W., ALLEN, M., MALAHLEHA, M., MNGADI, K. and DANIELS, B., 2018. Subtype C ALVAC-HIV and bivalent subtype C gp120/MF59 HIV-1 vaccine in low-risk, HIV-uninfected, South African adults: a phase 1/2 trial. *The lancet HIV,* **5**(7), pp. e366-e378.

BELOUKAS, A., PSARRIS, A., GIANNELOU, P., KOSTAKI, E., HATZAKIS, A. and PARASKEVIS, D., 2016. Molecular epidemiology of HIV-1 infection in Europe: An overview. *Infection, Genetics and Evolution,* **46**, pp. 180-189.

BENNETT, D.E., CAMACHO, R.J., OTELEA, D., KURITZKES, D.R., FLEURY, H., KIUCHI, M., HENEINE, W., KANTOR, R., JORDAN, M.R. and SCHAPIRO, J.M., 2009. Drug resistance mutations for surveillance of transmitted HIV-1 drug-resistance: 2009 update. *PloS one,* **4**(3), pp. e4724.

BISHNU, S., BANDYOPADHYAY, D., SAMUI, S., DAS, I., MONDAL, P., GHOSH, P., ROY, D. and MANNA, S., 2014. Assessment of clinico-immunological profile of newly diagnosed HIV patients presenting to a teaching hospital of eastern India. *The Indian journal of medical research,* **139**(6), pp. 903.

BLOOD, G.A.C., 2016. Human immunodeficiency virus (HIV). *Transfusion Medicine and Hemotherapy,* **43**(3), pp. 203.

BULL, J.J., SANJUAN, R. and WILKE, C.O., 2007. Theory of lethal mutagenesis for viruses. *Journal of virology*, **81**(6), pp. 2930-2939.

BUONAGURO, L., TORNESELLO, M. and BUONAGURO, F., 2007. Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. *Journal of virology,* **81**(19), pp. 10209-10219.

BURRELL, C.J., HOWARD, C.R. and MURPHY, F.A., 2017. Epidemiology of viral infections. *Fenner and White's Medical Virology,* , pp. 185.

CADIMA-COUTO, I. and GONCALVES, J., 2010. Towards inhibition of Vif-APOBEC3G interaction: which protein to target? *Advances in Virology*, **2010**.

CARR, J., WILLIAMS, D.G. and HAYDEN, R.T., 2010. Molecular detection of multiple respiratory viruses. *Molecular diagnostics.* Elsevier, pp. 289-300.

CASE, K., 1986. Nomenclature: human immunodeficiency virus. *Annals of Internal Medicine,* **105**(1), pp. 133-133.

CENTERS FOR DISEASE CONTROL AND PREVENTION, 2022-last update, **HIV Transmission | HIV Basics | HIV/AIDS | CDC** [Homepage of CDC], [Online]. Available: https://www.cdc.gov/hiv/basics/transmission.html [03/30, 2022].

CHAN, D.C. and KIM, P.S., 1998. HIV entry and its inhibition. *Cell,* **93**(5), pp. 681-684.

CHAPLIN, B., EISEN, G., IDOKO, J., ONWUJEKWE, D., IDIGBE, E., ADEWOLE, I., GASHAU, W., MELONI, S., SARR, A. and SANKALÉ, J., 2011. Impact of HIV type 1 subtype on drug resistance mutations in Nigerian patients failing first-line therapy. *AIDS Research and Human Retroviruses,* **27**(1), pp. 71-80.

CHEN, J., NIKOLAITCHIK, O., SINGH, J., WRIGHT, A., BENCSICS, C.E., COFFIN, J.M., NI, N., LOCKETT, S., PATHAK, V.K. and HU, W., 2009. High efficiency of HIV-1 genomic RNA packaging and heterozygote formation revealed by single virion analysis. *Proceedings of the National Academy of Sciences,* **106**(32), pp. 13535-13540.

CHRYSOSTOMOU, A.C., TOPCU, C., STYLIANOU, D.C., HEZKA, J. and KOSTRIKIS, L.G., 2020. Development of a new comprehensive HIV-1 genotypic drug resistance assay for all commercially available reverse transcriptase, protease and integrase inhibitors in patients infected with group M HIV-1 strains. *Infection, Genetics and Evolution,* **81**, pp. 104243.

CLAPHAM, P.R. and MCKNIGHT, A., 2001. HIV-1 receptors and cell tropism. *British medical bulletin,* **58**(1), pp. 43-59.

COAKLEY, E., PETROPOULOS, C.J. and WHITCOMB, J.M., 2005. Assessing chemokine co-receptor usage in HIV. *Current opinion in infectious diseases,* **18**(1), pp. 9-15.

COMPENDIUM, H.S., 2018. Foley B, LT, Apetrei C, Hahn B, Mizrachi I, Mullins J, Rambaut A, Wolinsky S & Korber B, Eds. *Published by Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, LA-UR,* **18**(25673), pp. 2018.

DALGLEISH, A.G., BEVERLEY, P.C., CLAPHAM, P.R., CRAWFORD, D.H., GREAVES, M.F. and WEISS, R.A., 1984. The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature,* **312**(5996), pp. 763-767.

DENG, H., LIU, R., ELLMEIER, W., CHOE, S., UNUTMAZ, D., BURKHART, M., MARZIO, P.D., MARMON, S., SUTTON, R.E. and HILL, C.M., 1996. Identification of a major co-receptor for primary isolates of HIV-1. *Nature,* **381**(6584), pp. 661-666.

DEREBAIL, S.S. and DESTEFANO, J.J., 2004. Mechanistic analysis of pause site-dependent and-independent recombinogenic strand transfer from structurally diverse regions of the HIV genome. *Journal of Biological Chemistry,* **279**(46), pp. 47446-47454.

DESTEFANO, J.J., BAMBARA, R.A. and FAY, P.J., 1994. The mechanism of human immunodeficiency virus reverse transcriptase-catalyzed strand transfer from internal regions of heteropolymeric RNA templates. *Journal of Biological Chemistry,* **269**(1), pp. 161-168.

DOMINGO, E., ESCARMÍS, C., MENÉNDEZ-ARIAS, L., PERALES, C., HERRERA, M., NOVELLA, I.S. and HOLLAND, J.J., 2008. Viral quasispecies: dynamics, interactions, and pathogenesis. *Origin and evolution of viruses,* , pp. 87-118.

DRAGIC, T., LITWIN, V., ALLAWAY, G.P., MARTIN, S.R., HUANG, Y., NAGASHIMA, K.A., CAYANAN, C., MADDON, P.J., KOUP, R.A. and MOORE, J.P., 1996. HIV-1 entry into CD4 cells is mediated by the chemokine receptor CC-CKR-5. *Nature,* **381**(6584), pp. 667-673.

FARIA, N.R., RAMBAUT, A., SUCHARD, M.A., BAELE, G., BEDFORD, T., WARD, M.J., TATEM, A.J., SOUSA, J.D., ARINAMINPATHY, N. and PÉPIN, J., 2014. The early spread and epidemic ignition of HIV-1 in human populations. *Science,* **346**(6205), pp. 56-61.

FENG, D. and DOOLITTLE, R.F., 1987. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal of Molecular Evolution,* **25**(4), pp. 351-360.

FENG, Y., ZHANG, C., ZHANG, M., GAO, L., MIAO, J., JIA, Y., DONG, X. and XIA, X., 2018. First report of a novel HIV-1 recombinant form (CRF100_01C) comprising CRF01_AE and C among heterosexuals in Yunnan, China. *Journal of Infection,* **77**(6), pp. 561-571.

FENG, Y., BRODER, C.C., KENNEDY, P.E. and BERGER, E.A., 1996. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science (New York, N.Y.),* **272**(5263), pp. 872-877.

FENYO, E.M., ALBERT, J. and ASJO, B., 1989. Replicative capacity, cytopathic effect and cell tropism of HIV. *AIDS (London, England),* **3 Suppl 1**, pp. S5-12.

FENYO, E.M., MORFELDT-MANSON, L., CHIODI, F., LIND, B., VON GEGERFELT, A., ALBERT, J., OLAUSSON, E. and ASJO, B., 1988. Distinct replicative and cytopathic characteristics of human immunodeficiency virus isolates. *Journal of virology,* **62**(11), pp. 4414-4419.

FOX, J. and FIDLER, S., 2010. Sexual transmission of HIV-1. *Antiviral Research,* **85**(1), pp. 276-285.

FRIEDMAN-KIEN, A.E., 1981. Disseminated Kaposi's sarcoma syndrome in young homosexual men. *Journal of the American Academy of Dermatology,* **5**(4), pp. 468-471.

GALAI, N. and KALINKOVICH, A., 1997. African HIV-1 subtype C and rate of progression among Ethiopian immigrants in Israel. *The Lancet,* **349**(9046), pp. 180-181.

GALE, C.V., MYERS, R., TEDDER, R.S., WILLIAMS, I.G. and KELLAM, P., 2004. Development of a novel human immunodeficiency virus type 1 subtyping tool, Subtype Analyzer (STAR): analysis of subtype distribution in London. *AIDS Research and Human Retroviruses,* **20**(5), pp. 457-464.

GALETTO, R., GIACOMONI, V., VÉRON, M. and NEGRONI, M., 2006. Dissection of a circumscribed recombination hot spot in HIV-1 after a single infectious cycle. *Journal of Biological Chemistry,* **281**(5), pp. 2711-2720.

GALLO, R.C., SARIN, P.S., GELMANN, E., ROBERT-GUROFF, M., RICHARDSON, E., KALYANARAMAN, V., MANN, D., SIDHU, G.D., STAHL, R.E. and ZOLLA-PAZNER, S., 1983. Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science,* **220**(4599), pp. 865-867.

GANSER-PORNILLOS, B.K., YEAGER, M. and SUNDQUIST, W.I., 2008. The structural biology of HIV assembly. *Current opinion in structural biology,* **18**(2), pp. 203-217.

GAO, F., BAILES, E., ROBERTSON, D.L., CHEN, Y., RODENBURG, C.M., MICHAEL, S.F., CUMMINS, L.B., ARTHUR, L.O., PEETERS, M. and SHAW, G.M., 1999. Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes. *Nature,* **397**(6718), pp. 436-441.

GASCHEN, B., TAYLOR, J., YUSIM, K., FOLEY, B., GAO, F., LANG, D., NOVITSKY, V., HAYNES, B., HAHN, B.H. and BHATTACHARYA, T., 2002. Diversity considerations in HIV-1 vaccine selection. *Science,* **296**(5577), pp. 2354-2360.

GÖTTE, M., LI, X. and WAINBERG, M.A., 1999. HIV-1 reverse transcription: A brief overview focused on structure–function relationships among molecules involved in initiation of the reaction. *Archives of Biochemistry and Biophysics,* **365**(2), pp. 199-210.

GREEN, M.R. and SAMBROOK, J., 2018. Touchdown Polymerase Chain Reaction (PCR). *Cold Spring Harbor protocols,* **2018**(5), pp. 10.1101/pdb.prot095133.

GUINDON, S., DUFAYARD, J., LEFORT, V., ANISIMOVA, M., HORDIJK, W. and GASCUEL, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology,* **59**(3), pp. 307-321.

HAASE, A.T., 1986. Pathogenesis of lentivirus infections. *Nature,* **322**(6075), pp. 130-136.

HEMELAAR, J., ELANGOVAN, R., YUN, J., DICKSON-TETTEH, L., FLEMINGER, I., KIRTLEY, S., WILLIAMS, B., GOUWS-WILLIAMS, E., GHYS, P.D. and ALASH'LE G, A., 2019. Global and regional molecular epidemiology of HIV-1, 1990–2015: a systematic review, global survey, and trend analysis. *The Lancet infectious diseases,* **19**(2), pp. 143-155.

HEMELAAR, J., ELANGOVAN, R., YUN, J., DICKSON-TETTEH, L., KIRTLEY, S., GOUWS-WILLIAMS, E., GHYS, P.D., ALASH'LE G, A., AGWALE, S. and ARCHIBALD, C., 2020. Global and regional epidemiology of HIV-1 recombinants in 1990–2015: A systematic review and global survey. *The Lancet HIV,* **7**(11), pp. e772-e781.

HO, D.D., NEUMANN, A.U., PERELSON, A.S., CHEN, W., LEONARD, J.M. and MARKOWITZ, M., 1995. Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. *Nature,* **373**(6510), pp. 123-126.

HOLEC, A.D., MANDAL, S., PRATHIPATI, P.K. and DESTACHE, C.J., 2017. Nucleotide reverse transcriptase inhibitors: a thorough review, present status and future perspective as HIV therapeutics. *Current HIV research,* **15**(6), pp. 411-421.

HU, W. and HUGHES, S.H., 2012. HIV-1 reverse transcription. *Cold Spring Harbor perspectives in medicine,* **2**(10), pp. a006882.

HU, W. and TEMIN, H.M., 1990. Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination. *Proceedings of the National Academy of Sciences,* **87**(4), pp. 1556-1560.

KALEEBU, P., FRENCH, N., MAHE, C., YIRRELL, D., WATERA, C., LYAGOBA, F., NAKIYINGI, J., RUTEBEMBERWA, A., MORGAN, D. and WEBER, J., 2002. Effect of human immunodeficiency virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1—positive persons in Uganda. *The Journal of infectious diseases,* **185**(9), pp. 1244-1250.

KANKI, P.J., HAMEL, D.J., SANKALÉ, J., HSIEH, C., THIOR, I., BARIN, F., WOODCOCK, S.A., GUÈYE-NDIAYE, A., ZHANG, E. and MONTANO, M., 1999. Human immunodeficiency virus type 1 subtypes differ in disease progression. *The Journal of infectious diseases,* **179**(1), pp. 68-73.

KATOH, K., ROZEWICKI, J. and YAMADA, K.D., 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in bioinformatics,* **20**(4), pp. 1160-1166.

KATOH, K. and STANDLEY, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution,* **30**(4), pp. 772-780.

KEARSE, M., MOIR, R., WILSON, A., STONES-HAVAS, S., CHEUNG, M., STURROCK, S., BUXTON, S., COOPER, A., MARKOWITZ, S. and DURAN, C., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics,* **28**(12), pp. 1647-1649.

KHER, U., 1982. A name for the plague: 80 days that changed the world. *Time. Retrieved from* http://content.time.com/time/specials/packages/article/0,28804,1977881_1977895_1978703,00.html.

KIRCHHOFF, F., 2013. HIV life cycle: overview. *Encyclopedia of AIDS*, , pp. 1-9.

KIWANUKA, N., ROBB, M., LAEYENDECKER, O., KIGOZI, G., WABWIRE-MANGEN, F., MAKUMBI, F.E., NALUGODA, F., KAGAAYI, J., ELLER, M. and ELLER, L.A., 2010. HIV-1 viral subtype differences in the rate of CD4 T-cell decline among HIV seroincident antiretroviral naive persons in Rakai district, Uganda. *Journal of acquired immune deficiency syndromes (1999)*, **54**(2), pp. 180.

KOMANDURI, K.V., VISWANATHAN, M.N., WIEDER, E.D., SCHMIDT, D.K., BREDT, B.M., JACOBSON, M.A. and MCCUNE, J.M., 1998. Restoration of cytomegalovirus-specific CD4 T-lymphocyte responses after ganciclovir and highly active antiretroviral therapy in individuals infected with HIV-1. *Nature medicine*, **4**(8), pp. 953-956.

KONSTANTINOV, I., STEFANOV, Y., KOVALEVSKY, A., VORONIN, Y., GRISHANIN, K. and GRISHAEV, M., 08.09.2010, Human immunodeficiency virus (HIV) [Homepage of Nature Medicine Supplementary cover], [Online]. Available: https://visual-science.com/projects/hiv/illustrations/ [03/01, 2022].

KORBER, B., GASCHEN, B., YUSIM, K., THAKALLAPALLY, R., KESMIR, C. and DETOURS, V., 2001. Evolutionary and immunological implications of contemporary HIV-1 variation. *British medical bulletin*, **58**(1), pp. 19-42.

KOSTRIKIS, L.G., HEZKA, J., STYLIANOU, D.C., KOSTAKI, E., ANDREOU, M., KOUSIAPPA, I., PARASKEVIS, D. and DEMETRIADES, I., 2018. HIV-1 transmission networks across Cyprus (2010-2012). *Plos one*, **13**(4), pp. e0195660.

KOSTRIKIS, L.G., HUANG, Y., MOORE, J.P., WOLINSKY, S.M., ZHANG, L., GUO, Y., DEUTSCH, L., PHAIR, J., U NEUMANN, A. and D HO, D., 1998. A chemokine receptor CCR2 allele delays HIV-1 disease progression and is associated with a CCR5 promoter mutation. *Nature medicine*, **4**(3), pp. 350-353.

KOUSIAPPA, I., ACHILLEOS, C., HEZKA, J., LAZAROU, Y., OTHONOS, K., DEMETRIADES, I. and KOSTRIKIS, L.G., 2011. Molecular characterization of HIV type 1 strains from newly diagnosed patients in Cyprus (2007–2009) recovers multiple clades including unique recombinant strains and lack of transmitted drug resistance. *AIDS Research and Human Retroviruses*, **27**(11), pp. 1183-1199.

KOUSIAPPA, I., VAN DE VIJVER, DAVID AMC and KOSTRIKIS, L.G., 2009. Near full-length genetic analysis of HIV sequences derived from Cyprus: evidence of a highly polyphyletic and evolving infection. *AIDS Research and Human Retroviruses*, **25**(8), pp. 727-740.

KREISS, J., 1997. Breastfeeding and vertical transmission of HIV-1. *Acta paediatrica,* **86**(S421), pp. 113-117.

KUMAR, S., STECHER, G., LI, M., KNYAZ, C. and TAMURA, K., 2018. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular biology and evolution,* **35**(6), pp. 1547-1549.

LAM, T.T., HON, C. and TANG, J.W., 2010. Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Critical reviews in clinical laboratory sciences,* **47**(1), pp. 5-49.

LARSSON, A., 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics,* **30**(22), pp. 3276-3278.

LEDERMAN, M.M., CONNICK, E., LANDAY, A., KURITZKES, D.R., SPRITZLER, J., ST. CLAIR, M., KOTZIN, B.L., FOX, L., HEATH CHIOZZI, M. and LEONARD, J.M., 1998. Immunologic responses associated with 12 weeks of combination antiretroviral therapy consisting of zidovudine, lamivudine, and ritonavir: results of AIDS Clinical Trials Group Protocol 315. *Journal of Infectious Diseases,* **178**(1), pp. 70-79.

LEFKOWITZ, E.J., DEMPSEY, D.M., HENDRICKSON, R.C., ORTON, R.J., SIDDELL, S.G. and SMITH, D.B., 2018. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic acids research,* **46**(D1), pp. D708-D717.

LEWINSKI, M., BISGROVE, D., SHINN, P., CHEN, H., HOFFMANN, C., HANNENHALLI, S., VERDIN, E., BERRY, C., ECKER, J. and BUSHMAN, F., 2005. Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *Journal of virology,* **79**(11), pp. 6610-6619.

LI, G. and DE CLERCQ, E., 2016. HIV genome-wide protein associations: a review of 30 years of research. *Microbiology and Molecular Biology Reviews,* **80**(3), pp. 679-731.

LI, H., DOU, J., DING, L. and SPEARMAN, P., 2007. Myristoylation is required for human immunodeficiency virus type 1 Gag-Gag multimerization in mammalian cells. *Journal of virology,* **81**(23), pp. 12899-12910.

LI, X., NING, C., HE, X., YANG, Y., LI, F., XING, H., HONG, K., YANG, R. and SHAO, Y., 2013. Genome sequences of a novel HIV-1 circulating recombinant form (CRF61_BC) identified among heterosexuals in China. *Genome announcements,* **1**(3), pp. e00326-13.

LOEMBA, H., BRENNER, B., PARNIAK, M.A., MA'AYAN, S., SPIRA, B., MOISI, D., OLIVEIRA, M., DETORIO, M. and WAINBERG, M.A., 2002. Genetic divergence of human immunodeficiency virus type 1 Ethiopian clade C reverse transcriptase (RT) and rapid development of resistance against nonnucleoside inhibitors of RT. *Antimicrobial Agents and Chemotherapy,* **46**(7), pp. 2087-2094.

LOLE, K.S., BOLLINGER, R.C., PARANJAPE, R.S., GADKARI, D., KULKARNI, S.S., NOVAK, N.G., INGERSOLL, R., SHEPPARD, H.W. and RAY, S.C., 1999. Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination. *Journal of virology*, **73**(1), pp. 152-160.

MANSKY, L.M., 1996. Forward mutation rate of human immunodeficiency virus type 1 in a T lymphoid cell line. *AIDS Research and Human Retroviruses*, **12**(4), pp. 307-314.

MANSKY, L.M. and TEMIN, H.M., 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *Journal of virology*, **69**(8), pp. 5087-5094.

MARZETTA, C.A., LEE, S.S., WROBEL, S.J., SINGH, K.J., RUSSELL, N. and ESPARZA, J., 2010. The potential global market size and public health value of an HIV-1 vaccine in a complex global market. *Vaccine*, **28**(30), pp. 4786-4797.

MINH, B.Q., NGUYEN, M.A.T. and VON HAESELER, A., 2013. Ultrafast approximation for phylogenetic bootstrap. *Molecular biology and evolution*, **30**(5), pp. 1188-1195.

MITSUYA, H. and BRODER, S., 1986. Inhibition of the in vitro infectivity and cytopathic effect of human T-lymphotrophic virus type III/lymphadenopathy-associated virus (HTLV-III/LAV) by 2', 3'-dideoxynucleosides. *Proceedings of the National Academy of Sciences*, **83**(6), pp. 1911-1915.

MITSUYA, H., WEINHOLD, K.J., FURMAN, P.A., ST CLAIR, M.H., LEHRMAN, S.N., GALLO, R.C., BOLOGNESI, D., BARRY, D.W. and BRODER, S., 1985. 3'-Azido-3'-deoxythymidine (BW A509U): an antiviral agent that inhibits the infectivity and cytopathic effect of human T-lymphotropic virus type III/lymphadenopathy-associated virus in vitro. *Proceedings of the National Academy of Sciences*, **82**(20), pp. 7096-7100.

MOUMEN, A., POLOMACK, L., UNGE, T., VÉRON, M., BUC, H. and NEGRONI, M., 2003. Evidence for a Mechanism of Recombination during Reverse Transcription Dependent on the Structure of the Acceptor RNA* 210. *Journal of Biological Chemistry*, **278**(18), pp. 15973-15982.

NATIONAL INSTITUTE OF HEALTH, 2011-last update, GLOSSARY of HIV/AIDS-Related Terms 2021. Available: https://clinicalinfo.hiv.gov/sites/default/files/glossary/Glossary-English_HIVinfo.pdf [03/27, 2022].

NGUYEN, L., SCHMIDT, H.A., VON HAESELER, A. and MINH, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, **32**(1), pp. 268-274.

NYAMWEYA, S., HEGEDUS, A., JAYE, A., ROWLAND-JONES, S., FLANAGAN, K.L. and MACALLAN, D.C., 2013. Comparing HIV-1 and HIV-2 infection: Lessons for viral immunopathogenesis. *Reviews in medical virology*, **23**(4), pp. 221-240.

PALMER, S., ALAEUS, A., ALBERT, J. and COX, S., 1998. Drug susceptibility of subtypes A, B, C, D, and E human immunodeficiency virus type 1 primary isolates. *AIDS Research and Human Retroviruses,* **14**(2), pp. 157-162.

PILLER, S., CALY, L. and JANS, D.A., 2003. Nuclear import of the pre-integration complex (PIC): the Achilles heel of HIV? *Current Drug Targets,* **4**(5), pp. 409-429.

PINEDA-PEÑA, A., FARIA, N.R., IMBRECHTS, S., LIBIN, P., ABECASIS, A.B., DEFORCHE, K., GÓMEZ-LÓPEZ, A., CAMACHO, R.J., DE OLIVEIRA, T. and VANDAMME, A., 2013. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. *Infection, genetics and evolution,* **19**, pp. 337-348.

PINEDA-PEÑA, A., SCHROOTEN, Y., VINKEN, L., FERREIRA, F., LI, G., TROVÃO, N.S., KHOURI, R., DERDELINCKX, I., DE MUNTER, P. and KÜCHERER, C., 2014. Trends and predictors of transmitted drug resistance (TDR) and clusters with TDR in a local Belgian HIV-1 epidemic. *PLoS One,* **9**(7), pp. e101738.

PINEDA-PEÑA, A., THEYS, K., STYLIANOU, D.C., DEMETRIADES, I., ABECASIS, A.B. and KOSTRIKIS, L.G., 2018. HIV-1 Infection in Cyprus, the Eastern Mediterranean European frontier: a densely sampled transmission dynamics analysis from 1986 to 2012. *Scientific Reports,* **8**(1), pp. 1-14.

PRESTON, B.D., ALBERTSON, T.M. and HERR, A.J., 2010. DNA replication fidelity and cancer, *Seminars in cancer biology* 2010, Elsevier, pp. 281-293.

QUAN, Y., LIANG, C., BRENNER, B.G. and WAINBERG, M.A., 2009. Multidrug-resistant variants of HIV type 1 (HIV-1) can exist in cells as defective quasispecies and be rescued by superinfection with other defective HIV-1 variants. *The Journal of infectious diseases,* **200**(9), pp. 1479-1483.

RAGONNET-CRONIN, M., HODCROFT, E., HUÉ, S., FEARNHILL, E., DELPECH, V., BROWN, A.J.L. and LYCETT, S., 2013. Automated analysis of phylogenetic clusters. *BMC bioinformatics,* **14**(1), pp. 1-10.

REEVES, J.D. and DOMS, R.W., 2002. Human immunodeficiency virus type 2. *Journal of general virology,* **83**(6), pp. 1253-1265.

RERKS-NGARM, S., PITISUTTITHUM, P., NITAYAPHAN, S., KAEWKUNGWAL, J., CHIU, J., PARIS, R., PREMSRI, N., NAMWAT, C., DE SOUZA, M. and ADAMS, E., 2009. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *New England Journal of Medicine,* **361**(23), pp. 2209-2220.

RICKMAN, L.S., 1997. Pneumocystis pneumonia-Los Angeles-Centers for Disease Control. MMWR 1981; 30: 250-252. *AJIC: American Journal of Infection Control,* **1**(25), pp. 68.

ROBERTS, J.D., BEBENEK, K. and KUNKEL, T.A., 1988. The accuracy of reverse transcriptase from HIV-1. *Science,* **242**(4882), pp. 1171-1173.

ROBERTSON, D.L., ANDERSON, J., BRADAC, J., CARR, J., FOLEY, B., FUNKHOUSER, R., GAO, F., HAHN, B., KALISH, M. and KUIKEN, C., 2000. HIV-1 nomenclature proposal. *Science,* **288**(5463), pp. 55-55.

RODA, R.H., BALAKRISHNAN, M., KIM, J.K., ROQUES, B.P., FAY, P.J. and BAMBARA, R.A., 2002. Strand transfer occurs in retroviruses by a pause-initiated two-step mechanism. *Journal of Biological Chemistry,* **277**(49), pp. 46900-46911.

ROM, W.N. and MARKOWITZ, S.B., 2007. *Environmental and occupational medicine.* Lippincott Williams & Wilkins.

ROZANOV, M., PLIKAT, U., CHAPPEY, C., KOCHERGIN, A. and TATUSOVA, T., 2004. A web-based genotyping resource for viral sequences. *Nucleic acids research,* **32**(suppl_2), pp. W654-W659.

SANTORO, M.M. and PERNO, C.F., 2013. HIV-1 genetic variability and clinical implications. *International Scholarly Research Notices,* **2013**.

SAXENA, S.K. and CHITTI, S.V., 2016. Molecular Biology and Pathogenesis of Retroviruses. *Advances in Molecular Retrovirology.* IntechOpen, .

SCHULTZ, A., BULLA, I., ABDOU-CHEKARAOU, M., GORDIEN, E., MORGENSTERN, B., ZOULIM, F., DENY, P. and STANKE, M., 2012. jpHMM: recombination analysis in viruses with circular genomes such as the hepatitis B virus. *Nucleic acids research,* **40**(W1), pp. W193-W198.

SEILLIER-MOISEIWITSCH, F., MARGOLIN, B.H. and SWANSTROM, R., 1994. Genetic variability of the human immunodeficiency virus: statistical and biological issues. *Annual Review of Genetics,* **28**(1), pp. 559-596.

SHARP, P.M., BAILES, E., CHAUDHURI, R.R., RODENBURG, C.M., SANTIAGO, M.O. and HAHN, B.H., 2001. The origins of acquired immune deficiency syndrome viruses: where and when? *Philosophical Transactions of the Royal Society of London.Series B: Biological Sciences,* **356**(1410), pp. 867-876.

SHARP, P.M. and HAHN, B.H., 2011. Origins of HIV and the AIDS pandemic. *Cold Spring Harbor perspectives in medicine,* **1**(1), pp. a006841.

SIEPEL, A.C., HALPERN, A.L., MACKEN, C. and KORBER, B.T., 1995. A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Research and Human Retroviruses,* **11**(11), pp. 1413-1416.

SMYTH, R.P., DAVENPORT, M.P. and MAK, J., 2012. The origin of genetic diversity in HIV-1. *Virus research,* **169**(2), pp. 415-429.

SSEMWANGA, D., NSUBUGA, R.N., MAYANJA, B.N., LYAGOBA, F., MAGAMBO, B., YIRRELL, D., VAN DER PAAL, L., GROSSKURTH, H. and KALEEBU, P., 2013. Effect of HIV-1 subtypes on disease progression in rural Uganda: a prospective clinical cohort study. *PloS one,* **8**(8), pp. e71768.

SUNDQUIST, W.I. and KRAUSSLICH, H.G., 2012. HIV-1 assembly, budding, and maturation. *Cold Spring Harbor perspectives in medicine,* **2**(7), pp. a006924.

TANG, M.W., LIU, T.F. and SHAFER, R.W., 2012. The HIVdb system for HIV-1 genotypic resistance interpretation. *Intervirology,* **55**(2), pp. 98-101.

TEMIN, H.M., 1991. Sex and recombination in retroviruses. *Trends in Genetics,* **7**(3), pp. 71-74.

TERSMETTE, M., DE GOEDE, R.E., AL, B.J., WINKEL, I.N., GRUTERS, R.A., CUYPERS, H.T., HUISMAN, H.G. and MIEDEMA, F., 1988. Differential syncytium-inducing capacity of human immunodeficiency virus isolates: frequent detection of syncytium-inducing isolates in patients with acquired immunodeficiency syndrome (AIDS) and AIDS-related complex. *Journal of virology,* **62**(6), pp. 2026-2032.

THOMPSON, J.D., HIGGINS, D.G. and GIBSON, T.J., 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research,* **22**(22), pp. 4673-4680.

TRIFINOPOULOS, J., NGUYEN, L., VON HAESELER, A. and MINH, B.Q., 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic acids research,* **44**(W1), pp. W232-W235.

TURNER, B.G. and SUMMERS, M.F., 1999. Structural biology of HIV. *Journal of Molecular Biology,* **285**(1), pp. 1-32.

VALLARI, A., HOLZMAYER, V., HARRIS, B., YAMAGUCHI, J., NGANSOP, C., MAKAMCHE, F., MBANYA, D., KAPTUÉ, L., NDEMBI, N. and GÜRTLER, L., 2011. Confirmation of putative HIV-1 group P in Cameroon. *Journal of virology,* **85**(3), pp. 1403-1407.

VAN HEUVERSWYN, F., LI, Y., NEEL, C., BAILES, E., KEELE, B.F., LIU, W., LOUL, S., BUTEL, C., LIEGEOIS, F. and BIENVENUE, Y., 2006. SIV infection in wild gorillas. *Nature,* **444**(7116), pp. 164-164.

VASAN, A., RENJIFO, B., HERTZMARK, E., CHAPLIN, B., MSAMANGA, G., ESSEX, M., FAWZI, W. and HUNTER, D., 2006. Different rates of disease progression of HIV type 1 infection in Tanzania based on infecting subtype. *Clinical Infectious Diseases,* **42**(6), pp. 843-852.

VENNER, C.M., NANKYA, I., KYEYUNE, F., DEMERS, K., KWOK, C., CHEN, P., RWAMBUYA, S., MUNJOMA, M., CHIPATO, T. and BYAMUGISHA, J., 2016. Infecting HIV-1 subtype predicts disease progression in women of sub-Saharan Africa. *EBioMedicine,* **13**, pp. 305-314.

WADMAN, M. and YOU, J., 2017. *The vaccine wars, .*

WATTS, J.M., DANG, K.K., GORELICK, R.J., LEONARD, C.W., BESS JR, J.W., SWANSTROM, R., BURCH, C.L. and WEEKS, K.M., 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature,* **460**(7256), pp. 711-716.

WILEN, C.B., TILTON, J.C. and DOMS, R.W., 2012. HIV: cell binding and entry. *Cold Spring Harbor perspectives in medicine,* **2**(8), pp. a006866.

WILEY, E.O. and LIEBERMAN, B.S., 2011. *Phylogenetics: theory and practice of phylogenetic systematics.* John Wiley & Sons.

WU, W., BLUMBERG, B.M., FAY, P.J. and BAMBARA, R.A., 1995. Strand transfer mediated by human immunodeficiency virus reverse transcriptase in vitro is promoted by pausing and results in misincorporation. *Journal of Biological Chemistry,* **270**(1), pp. 325-332.

WU, Y. and MARSH, J.W., 2003. Gene transcription in HIV infection. *Microbes and Infection,* **5**(11), pp. 1023-1027.

WYMANT, C., BEZEMER, D., BLANQUART, F., FERRETTI, L., GALL, A., HALL, M., GOLUBCHIK, T., BAKKER, M., ONG, S.H. and ZHAO, L., 2022. A highly virulent variant of HIV-1 circulating in the Netherlands. *Science,* **375**(6580), pp. 540-545.

YAMAGUCHI, J., VALLARI, A., MCARTHUR, C., STHRESHLEY, L., CLOHERTY, G.A., BERG, M.G. and RODGERS, M.A., 2020. Brief report: complete genome sequence of CG-0018a-01 establishes HIV-1 subtype L. *Journal of acquired immune deficiency syndromes (1999),* **83**(3), pp. 319.

YASEEN, M.M., ABUHARFEIL, N.M., ALQUDAH, M.A. and YASEEN, M.M., 2017. Mechanisms and factors that drive extensive human immunodeficiency virus type-1 hypervariability: an overview. *Viral immunology,* **30**(10), pp. 708-726.

ZHANG, C., FENG, Y., GAO, L., ZHANG, M., MIAO, J., DONG, X. and XIA, X., 2019. Genetic characterization and recombinant history of a novel HIV-1 circulating recombinant form (CRF101_01B) identified in Yunnan, China. *Infection, Genetics and Evolution,* **73**, pp. 109-112.